



Création d'un environnement de gestion de base de données " en grille ". Application à l'échange de données médicales.

Paul De Vlieger

► To cite this version:

Paul De Vlieger. Création d'un environnement de gestion de base de données " en grille ". Application à l'échange de données médicales.. Calcul parallèle, distribué et partagé [cs.DC]. Université d'Auvergne - Clermont-Ferrand I, 2011. Français. <NNT : 2011CLF1MM11>. <tel-00654660>

HAL Id: tel-00654660

<https://tel.archives-ouvertes.fr/tel-00654660>

Submitted on 22 Dec 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITE D'Auvergne
2011

ECOLE DOCTORALE
DES SCIENCES POUR L'INGENIEUR
N° d'ordre :

Thèse

Présentée à l'Université d'Auvergne
pour l'obtention du grade de DOCTEUR
(Décret du 5 juillet 1984)

Spécialité Informatique

soutenue le 12 Juillet 2011

DE VLIEGER Paul

Création d'un environnement de gestion de base de
données « en grille ». Application à l'échange de
données médicales

M. JEAN-YVES BOIRE
M. VINCENT BRETON
M. TRISTAN GLATARD
M. MICHEL DAYDE
M^{lle} LYDIA MAIGNE
M. JOHAN MONTAGNAT
M. PASCAL STACCINI

Directeur de Thèse
Directeur de Thèse
Examineur
Examineur
Examineur
Rapporteur
Rapporteur

Professeur à l'Université d'Auvergne
Directeur de Recherche au CNRS, Clermont-Ferrand
Chargé de Recherche au CNRS, CREATIS UMR 5520
Professeur à l'Université de Toulouse
Maître de conférences à l'Université Blaise Pascal
Directeur de Recherche au CNRS, Nice Sophia Antipolis
Professeur à l'Université de Nice Sophia Antipolis

ISIT

Laboratoire de Physique Corpusculaire

À mon épouse, Amélie

Remerciements

En premier lieu je tiens à remercier les membres du jury, tout d'abord JOHAN MONTAGNAT, *directeur de recherche CNRS à Polytech' Nice - Sophia Antipolis* et PASCAL STACCINI, *Professeur à l'université de Nice - Sophia Antipolis* pour m'avoir fait l'honneur d'être rapporteurs de ce manuscrit. Merci à MICHEL DAYDÉ, *Professeur à l'université de Toulouse*, TRISTAN GLATARD, *chargé de recherche CNRS à l'INSA de Lyon* et LYDIA MAIGNE, *maître de conférences à l'Université Blaise Pascal* d'avoir accepté d'être examinateurs de cette thèse.

J'adresse mes plus vifs remerciements à mes directeurs de thèse, JEAN-YVES BOIRE et VINCENT BRETON qui m'ont ouvert le chemin de la recherche et m'ont accompagné dans son exploration. Les connaissances, compétences et l'expertise de Jean-Yves sur le plan médical au sein de l'ERIM a naturellement amené l'originalité de ce sujet de cette thèse. L'apport technique et méthodologique de Vincent dans l'équipe PCSV du LPC ainsi que l'ouverture sur le monde des grilles a amené la complémentarité nécessaire à l'avènement de ce travail. Je lui loue ses qualités de pédagogue, d'écoute, d'attention et de rigueur scientifique qui restera un modèle pour moi. Je remercie aussi la région Auvergne qui a permis de réaliser financièrement ces travaux.

Je remercie l'ensemble des personnes qui ont contribué à l'élaboration du cahier des charges du projet *Réseau Sentinelle Cancer Auvergne (RSCA)* dont la liste est trop longue pour figurer ici. Je tiens cependant à adresser ma profonde reconnaissance à ANDRÉ LAUTIER, *président de l'association RSCA*, pour son avidité à servir le combat contre le cancer, pour avoir mené un grand nombre de discussions et pris de nombreuses décisions pour le projet, ce fut une source permanente de motivation pour moi. Dans le même registre, j'adresse un grand merci à CHANTAL MESTRE pour sa disponibilité permanente au sein de l'ARDOC et bénévolement pour RSCA. Merci aussi au DR. ALAIN GAILLOT, *président de l'ARDOC* pour nous avoir expliqué toutes les ficelles de l'anatomopathologie en Auvergne et pour nous avoir ouvert les portes des structures de dépistage organisé des cancers et de la SIPATH lors de la mise en œuvre de RSCA.

Je souhaite aussi remercier les membres de l'ERIM/ISIT pour m'avoir supporté au quotidien et pour les nombreux échanges que j'ai pu avoir avec eux : Laurent(s), Souha, Jérôme, Lemlih, Aline, Sylvie, Florian et bien d'autres. Il en est de même pour l'équipe PSCV du LPC : les nombreux échanges que j'ai pu avoir avec Matteo et Simon lors du développement de HOPE ont été extrêmement précieux pour l'architecture de RSCA. Le support technique et mental que m'ont offert Jean et Vincent pour gérer les caprices de la grille m'a permis par moments un gain de temps considérable. Merci aussi à Géraldine et son éternelle bonne humeur. Enfin merci à tous les autres membres anciens comme nouveaux de l'équipe pour tous les bons moments passés.

Je tiens à remercier chaleureusement Lydia, pour tous les échanges que nous avons eus et pour son rôle dans RSCA et ma thèse. Son aide fut très précieuse pour le franchissement de nombreuses étapes techniques lors de la mise en œuvre du projet. Sa disponibilité malgré ses nouvelles responsabilités dans PSCV et son support dans ces derniers mois a été extrêmement importante. Je lui adresse par ailleurs tous mes vœux de réussite pour le projet ANR GINSENG.

Je souhaite aussi exprimer ma gratitude envers Jonathan, ma première expérience d'encadrement de stage, pour le travail de très bonne qualité portant sur le sujet ingrat du dépoussiérage de l'environnement d'authentification CPS. Merci aussi à Charlotte pour avoir proposé les bases des outils de *Record Linkage* évoqués dans ce manuscrit.

Merci aussi à mes nouveaux collègues de l'IUT d'Allier, qui m'ont offert d'excellentes conditions de travail pour terminer la rédaction de ce manuscrit.

Je finirai ces remerciements sur une note plus personnelle, avec une pensée pour ma famille, source de conseil, d'inspiration, de réconfort et de bonheur permanent malgré l'éloignement. Merci aussi à Jo et Marcel pour m'avoir fait tant aimer les Auvergnats. Enfin comment ne pourrais-je pas remercier celle qui m'a toujours soutenu, qui a supporté mes humeurs et avec qui j'ai partagé tant de moments magiques : merci à toi Amélie.

Sommaire

INTRODUCTION GENERALE.....	13
CHAPITRE 1. LES ENJEUX D'UN SYSTEME DISTRIBUE DE GESTION DE BASES DE DONNEES POUR LA SANTE	17
INTRODUCTION	17
1.1. L'E-SANTE EN FRANCE EN 2010	18
1.1.1. L'organisation du système de soin français	18
1.1.2. L'organisation des systèmes d'information médicaux.....	18
1.1.3. L'e-santé.....	19
1.1.4. Dispositions légales pour le partage de données électroniques de santé.....	20
1.2. ORGANISATION DU DEPISTAGE ORGANISE DES CANCERS EN FRANCE	24
1.2.1. Le cancer	24
1.2.2. Les sources de données sur le cancer.....	26
1.2.3. L'épidémiologie et la veille sanitaire.....	28
1.2.4. Le dépistage	29
1.2.5. Efficacité du dépistage organisé	33
1.3. APPLICATION DES GRILLES POUR LA SANTE.....	33
1.3.1. Evolution de la gestion de la donnée médicale.....	33
1.3.2. Les grilles informatiques	34
1.3.3. Applications des grilles.....	37
1.3.4. Les différents types de grilles informatiques.....	41
1.3.5. Les grilles et la protection de la vie privée	45
1.3.6. Transposition pour le dépistage des cancers en Auvergne	46
CONCLUSION	49
CHAPITRE 2. CAHIER DES CHARGES DU PROJET RSCA.....	51
INTRODUCTION	51
1 INTRODUCTION.....	5
1.1 Principes fondateurs	5
1.2 Description du projet.....	6
1.3 Intérêt et objectifs généraux du projet	7
2 ACTEURS.....	8
2.1 Les acteurs/partenaires concernés	8
2.2 Les acteurs du projet "réseau sentinelle"	9
3 ANALYSE.....	13
3.1 Fonctionnement actuel	13

3.2	Analyse de l'existant	14
3.3	Objectif.....	15
3.4	Perspectives (objectif)	17
3.5	Moyens du projet	19
3.6	Bénéfices attendus.....	19
3.7	Liste des objectifs :	20
3.8	Perspectives de solution.....	22
4	DESCRIPTION FONCTIONNELLE	23
4.1	Caractéristiques et fonctionnement du système.....	23
4.2	Scénarios d'utilisation du système par acteur	24
5	DONNEES.....	28
5.1	Types de données.....	28
5.2	Données par Acteur.....	28
6	SECURITE	31
6.1	Identification utilisateur.....	31
6.2	Identification patient	32
7	ECHEANCIERS (MISE EN ŒUVRE).....	33
8	VALIDATION	34
9	ANNEXES.....	35
9.1	Bible des données anatomo-pathologiques standardisée	35
9.2	Glossaire (vocabulaire).....	36
10	CREDITS.....	37
10.1	Acteurs techniques.....	37
10.2	Autres acteurs	37
10.3	Anatomo-pathologistes	37
10.4	Associations	38
10.5	Epidémiologie.....	38
	CONCLUSION	91

CHAPITRE 3. MISE EN ŒUVRE DU RESEAU SENTINELLE CANCER AUVERGNE ..

INTRODUCTION	93
3.1. LES SYSTEMES ET STANDARDS DE GESTION DE L'INFORMATION DE SANTE	94
3.1.1. Les systèmes d'information et de communication de la santé	94
3.1.2. Comparaison avec les systèmes existants.....	95
3.1.3. Standards des dossiers médicaux partagés	99
3.2. LES TECHNOLOGIES INNOVANTES DE GRILLES POUR RSCA.....	102
3.2.1. La grille EGEE-EGI.....	102
3.2.2. L'intergiciel gLite adapté à RSCA.....	105
3.2.3. Mise en œuvre des concepts de grille pour RSCA.....	117
3.3. L'ARCHITECTURE INFORMATIQUE POUR LE RESEAU SENTINELLE CANCER AUVERGNE	118
3.3.1. Mise en œuvre de l'architecture	118
3.3.2. L'architecture choisie	128
3.4. MISE EN APPLICATION POUR LE DEPISTAGE ORGANISE DES CANCERS DU SEIN ET DU COLON	131
3.4.1. La structuration des données médicales.....	131
3.4.2. L'interfaçage avec les bases de données de pathologie	133
3.4.3. L'interfaçage avec les systèmes d'information métier.....	134

3.4.4. Bilan de mise en place du réseau	135
CONCLUSION	137
CHAPITRE 4. GESTION DU PATIENT ET DES DONNEES MEDICALES POUR RSCA	139
INTRODUCTION	139
4.1. SECURITE DES DONNEES MEDICALES.....	141
4.1.1. Evaluation des risques.....	141
4.1.2. Eléments de sécurité requis par la CNIL	144
4.1.3. Utilisation de la Carte de Professionnel de Santé (CPS).....	144
4.2. L'IDENTIFICATION DU PATIENT	149
4.2.1. Les enjeux de l'identification.....	149
4.2.2. Contraintes d'un système d'identification	151
4.3. MODELE D'IDENTIFICATION DYNAMIQUE ET DISTRIBUE DU PATIENT.....	152
4.3.1. Présentation de la solution adoptée	152
4.3.2. Description des scénarios d'identification.....	154
4.4. RAPPROCHEMENT D'IDENTITES MEDICALES DISTRIBUEES	158
4.4.1. Enjeu du rapprochement des patients pour l'épidémiologie	159
4.4.2. Les techniques de rapprochement de données.....	159
4.4.3. Techniques de comparaison.....	160
4.4.4. Medical Data Linkage	165
4.5. PRESENTATION DES RESULTATS	169
4.5.1. Expérimentation sur données simulées.....	169
4.5.2. Expérimentation sur données réelles.....	171
4.5.3. Expérimentation étendue sur données réelles.....	172
CONCLUSION ET PERSPECTIVES	175
CONCLUSION GENERALE.....	177
ANNEXES.....	180
Annexe 1. Modèle de données Anatomo-pathologiques	181
Annexe 2. Interface d'authentification Gateway.....	185
Annexe 3. Interface Zeus d'OSI-Santé pour les associations de dépistage	186
Annexe 4. L'algorithme Phonex [198]	187
Annexe 5. Comparaison des algorithmes de « data linkage ».....	188
Annexe 6. Version GP-GPU de Jaro-Winkler.....	189
BIBLIOGRAPHIE	193
LISTE DES PUBLICATIONS	200
TABLE DES FIGURES.....	201
RESUME	204

Introduction générale

L'explosion du volume de données produites par les activités humaines a fortement influencé l'évolution des systèmes de gestion de bases de données. Leurs capacités, performances et fiabilité se sont fortement accrues pour pouvoir absorber en toute sécurité des quantités toujours plus importantes d'informations.

La multiplication des sources de données a par ailleurs amené une grande dispersion de l'information personnelle. Les regrouper pose de nombreux problèmes d'origines différentes, que ce soit d'ordre technique ou légal. Du point de vue technologique, l'évolution des bases de données, focalisée en grande partie sur l'amélioration des moteurs a cependant délaissé certains aspects liés à l'interconnexion, l'accès distant et l'interopérabilité. Du point de vue légal, le regroupement d'informations personnelles est soumis à de nombreuses contraintes juridiques fixées par les différentes législations.

Le contexte médical est au carrefour de toutes ces problématiques, avec un volume de données gigantesque produit chaque jour qui, de plus, est considérablement dispersé dans les différents systèmes d'information des établissements de santé. Le regroupement de ces informations est au centre de nombreux projets de constitution du « dossier médical partagé » dont la lenteur de mise en œuvre et les coûts de développement confirment sa difficulté.

Le dépistage organisé des cancers n'échappe pas au problème et de réelles difficultés d'accès aux données sont rencontrées au quotidien par les structures organisant le dépistage. N'ayant ni le droit, ni les moyens d'accéder de façon informatisée aux comptes rendus médicaux, la collecte de données s'effectue par des moyens non-numériques qui sont lourds de conséquences sur le plan humain et financier.

L'accès distant, protégé et sécurisé à des masses de données distribuées est pourtant une problématique largement éprouvée dans le monde de la recherche. Le besoin de stockage et d'analyse des expériences de collision de particules réalisées dans les accélérateurs tels que le LHC au CERN a été à la genèse des technologies de grilles informatiques. Ces outils permettent de mettre en relation grâce aux capacités des réseaux informatiques un grand nombre d'ordinateurs en mutualisant leurs capacités de calcul et de stockage.

Le but du travail présenté dans ce mémoire est la mise en œuvre d'une architecture s'inspirant des grilles informatiques dans le cadre du dépistage organisé des cancers en Auvergne. L'infrastructure ainsi créée porte le nom de « Réseau Sentinelle Cancer Auvergne » ou *RSCA*.

Le premier objectif est de fournir une analyse des besoins et contraintes que posent l'échange de données médicales avec une étude spécifique au cas d'utilisation du dépistage organisé des cancers. Une analyse de la capacité des technologies des grilles informatiques à répondre à ce défi sera de mise, mais celle-ci ne peut se faire sans connaissance du cadre juridique applicable en France et en Europe.

Ensuite, un cahier des charges est présenté, en reprenant l'ensemble des objectifs fonctionnels, techniques et juridiques en accord avec l'ensemble des acteurs du dépistage des cancers en Auvergne et de leurs contraintes fonctionnelles. Une architecture répondant à toutes ces exigences est proposée en réutilisant les capacités des grilles informatiques à pouvoir fédérer un ensemble de sources de données distribuées.

Un ensemble de verrous scientifiques seront levés concernant des techniques, outils et méthodologies manquants à la mise en relation de bases de données médicales. Ceci sera abordé par l'étude de la problématique centrale de l'identification du patient au sein de sources de données distribuées et aux techniques de mise en relation et de dédoublement de l'information médicale. Ces étapes sont primordiales à l'exploitation des données, que ce soit d'un point de vue microscopique (transport d'un compte-rendu médical) ou macroscopique (analyse épidémiologique à grande échelle).

Ce mémoire est présenté en quatre parties. Le premier chapitre vise à présenter le cadre global de la santé en France, l'organisation du système de soin en général et le contexte du cancer par la suite. Le périmètre légal est aussi évoqué, permettant enfin de présenter en quoi les grilles informatiques peuvent répondre au besoin du dépistage des cancers.

Le deuxième chapitre présente le cahier des charges du projet RSCA : *Réseau Sentinelle Cancer Auvergne*, établi en début de thèse qui a été adopté par le consortium RSCA. Cette partie indique les différents acteurs régionaux et leurs besoins, à savoir principalement les structures de pathologie fournissant les rapports médicaux aux associations de dépistage organisé des cancers ainsi que les établissements régionaux de santé publique. L'ensemble des besoins et des contraintes du projet sont fixées, un modèle des données à traiter est proposé et les différents cas d'utilisation de l'application sont indiqués.

Le troisième chapitre constitue la réponse technique au cahier des charges en conformité avec les différents standards de la communication du monde médical. Il permet de présenter l'ensemble des outils nécessaires à la réalisation du projet pour proposer enfin l'architecture qui sera adoptée. Les technologies issues du monde des grilles informatiques pouvant faire l'objet d'une réutilisation ou d'une adaptation sont aussi présentées. Une mise en évidence est ensuite faite sur les outils qui restent encore à développer pour satisfaire l'ensemble des contraintes du cahier des charges.

Le point central d'un système distribué de bases de données est d'assurer sa cohérence, en garantissant qu'un patient soit identifié correctement au sein du réseau ainsi créé. Pour des raisons légales, il n'est pas possible d'utiliser comme identifiant le numéro de sécurité sociale en France. Il est alors nécessaire d'avoir recours à d'autres moyens pour rapprocher les patients situés dans différentes bases.

Ainsi, le dernier chapitre fait l'objet de ces développements effectués spécifiquement pour le *Réseau Sentinelle Cancer Auvergne*.

Un modèle de système d'identification respectant la sécurité et la confidentialité des données médicales est d'abord proposé, en présentant un ensemble de solutions permettant à la fois une certaine souplesse dans les possibilités de manipulation dynamique d'identités tout en assurant que la vie privée du patient est respectée.

La dernière étape consiste alors à rapprocher les patients identiques au sein du réseau sous une même identité en les comparant de façon empirique, c'est-à-dire mesurant trait par trait leur similarité. Cette discipline scientifique appelée « *record linkage* », propose des méthodes de comparaison qui seront étudiées afin de proposer un algorithme de rapprochement d'identités le plus robuste vis-à-vis des erreurs de retranscription que peut comporter une base de données.

Chapitre 1. Les enjeux d'un système distribué de gestion de bases de données pour la santé

INTRODUCTION

Parallèlement à la généralisation de l'informatisation des systèmes d'information pour la gestion des données de santé, l'accès à l'information depuis ces différents systèmes comporte des enjeux majeurs pour la santé. En effet, cette multiplication de sources de données médicales a fait naturellement naître de nouveaux espoirs en termes d'utilisation de cette masse d'information.

D'un point de vue médical, il paraissait évident que l'informatisation allait faciliter non seulement la gestion des données mais aussi leurs possibilités d'exploitation, que ce soit en clinique ou en santé publique.

D'un point de vue informatique, de nouvelles problématiques sont apparues pour permettre cette exploitation. Située au centre des enjeux éthiques, légaux, organisationnels et aussi techniques, l'informatique médicale se voit confrontée à de nombreux défis lorsqu'il s'agit d'agréger différentes sources d'informations.

Par ailleurs, la récupération puis la concentration de l'information en un immense dépôt centralisé de données a montré ses limites, que ce soit en termes d'efficacité ou en termes de coût. De ces constatations est née une toute autre méthodologie qui, à l'inverse d'un système centralisé va se connecter aux systèmes existants pour recueillir l'information nécessaire. Outre le coût moindre d'une telle solution, la disponibilité des données est extrêmement rapide, ce qui permet de répondre à un éventail de questions en santé publique avec une réactivité inédite.

Cependant, un système décentralisé se heurte à d'autres défis pour se conformer aux contraintes légales du domaine médical, ce chapitre établit ces exigences avec une application au domaine du cancer et propose une piste technologique de réponse.

1.1. L'E-SANTE EN FRANCE EN 2010

1.1.1. *L'organisation du système de soin français*

Le fonctionnement de la santé en France suit approximativement le schéma organisationnel des institutions françaises. Il est administré à un échelon national directement par l'Etat, qui le finance en grande partie. Les principaux acteurs sont le ministère de la santé au premier abord mais aussi le ministère du travail et des comptes publics qui jouent un rôle dans le domaine de la santé publique. D'autres structures sont chargées de certaines missions comme :

- la Haute Autorité de Santé (HAS), chargée de la qualité et de la cohérence des soins ;
- le Haut Conseil de la santé publique (HCSP) ;
- l'Agence Française de Sécurité Sanitaire des Produits de Santé (AFSSAPS).

A l'échelle régionale, l'Etat délègue à des structures locales certaines responsabilités et missions. Ces acteurs régionaux sont principalement les Agences Régionales de Santé (ARS) qui regroupent un ensemble de compétences régionales de santé. Elles ont pour mission d'appliquer la politique nationale de santé à leur échelle. Leur proximité facilite les rapports et la coordination avec les acteurs de santé locaux. A l'échelle régionale, des commissions et conférences, placées sous la tutelle des ARS et composées d'un collège d'acteurs de santé, d'usagers ou d'associations sont consultées pour partager leur avis lors de la prise de décisions de santé.

1.1.2. *L'organisation des systèmes d'information médicaux*

Parallèlement au développement de la santé, les technologies de l'information et de la communication (TIC) ont naturellement trouvé leur place au sein des structures médicales dès les prémices de l'informatique au courant des années 50. Les applications au domaine médical sont alors de plusieurs natures :

- en médecine clinique, avec l'aide au soin, au diagnostic : développement de l'imagerie médicale et du traitement de l'information ;
- dans l'organisation médicale : gestion des données générées par les flux de patients : création de bases de données et de systèmes d'information dédiés, naissance des standards (DICOM [1], HL7 [2]) et des protocoles de communications (PACS¹), début des réflexions sur les dossiers patients électroniques (EHR²).

Cependant, on remarque que les deux pans de l'informatique médicale n'ont pas suivi la même tendance. Là où l'informatique « clinique » est à la pointe de la technologie et de la recherche, avec les algorithmes les plus évolués en traitement du signal et de l'image, l'informatique « organisationnelle » peine à suivre les évolutions d'autres secteurs privés, comme le bancaire ou les transports etc.

¹ PACS : Picture Archiving and Communication System

² EHR : Electronic Health Record

En effet, même si l'avènement des réseaux et l'agrandissement des capacités de stockage ont donné naissance au concept d'hôpitaux « sans papier », censé réduire considérablement les coûts globaux de la chaîne de soin et améliorer de façon générale la qualité, l'application du concept peine à s'instaurer. De plus, le contexte juridique extrêmement tendu en France sur la gestion de données personnelles, et de surcroît médicales, freine considérablement le développement de systèmes évolués de gestion de données et pousse au cloisonnement et à la centralisation.

Les différents programmes, GMSIH¹ [3] et PMSI² [4], qui tentent de moderniser les systèmes d'information hospitaliers français n'ont pas totalement réussi à résoudre ces différents problèmes d'accès à l'information médicale. Ils se heurtent toujours à des obstacles liés à la confidentialité, l'identification des patients comme du personnel médical ou l'interopérabilité des systèmes, ne permettant pas d'aboutir à un dossier médical électronique qui pourtant constitue un enjeu majeur du système de soin moderne. Ainsi, la continuité des soins inter-établissement n'est toujours pas d'actualité, tout comme l'accès à l'historique d'un patient, et cela même pour une situation d'urgence. Il n'y a pas que le patient qui est lésé mais tout le système médical. Par conséquent, toute étude épidémiologique peine à élaborer des analyses statistiques précises par absence de disponibilité ou manque de fiabilité des données.

Plus récemment, le projet de loi de financement de la sécurité sociale a prévu la création de l'ASIP³ santé [5] et son portail *esante.gouv.fr*. Cette structure vise à rassembler en une même entité les anciens groupements d'intérêt public du gip-cps⁴ et du gip-dmp⁵. Ses missions sont de fournir un ensemble de spécifications pour le projet de dossier médical partagé français. Au programme se trouvent aussi la création d'un identifiant national de santé (INS), la gestion des cartes de professionnel de santé, la promotion de la télémédecine ou encore la supervision des agréments d'hébergeurs de données de santé. Ainsi, bon nombre de données médicales vont être traitées et gérées par ces consortiums privés. Ces éléments posent aussi le problème de la propriété des données, qui, en temps normal doivent toujours être accessibles par le patient. A ce propos, il est aussi extrêmement difficile de proposer au patient de consulter et surtout modifier les données le concernant, comme le stipule la loi informatique et libertés [1.1.4.2]. De plus, en cas de modification, les moyens d'avertir le médecin qui accède au dossier médical ne sont absolument pas pris en compte.

1.1.3. L'e-santé

Le préfixe “e” et le symbole “-” représentent le plus souvent l'interface entre l'informatique (IT) et les télécommunications (ICT). Ce préfixe est maintenant étendu à tout un ensemble de domaines, regroupé par le terme e-technologie comme e-mail, e-business, e-marketing, e-learning, e-engineering et même e-science... Avec bien entendu ses dérivations en langue française : e-commerce, e-administration etc.

¹ Groupement pour la Modernisation du Système d'Information Hospitalier

² Programme de Médicalisation des Systèmes d'Information

³ Agence des Systèmes d'Information de santé Partagé

⁴ Groupement d'Intérêt Public – Carte de Professionnel de Santé

⁵ Groupement d'Intérêt Public – Dossier Médical Partagé

Le domaine médical n'a pas échappé à la règle et le concept d'e-health ou d'e-santé est naturellement apparu. Il représente une ouverture de la santé aux moyens de communication. Les disciplines tout comme le public concerné sont vastes, les moyens et technologies mis en œuvre sont extrêmement variés. L'essor de l'e-santé en France est considérable et largement soutenu par la commission Européenne. Au vu des enjeux financiers sous-jacents, les moyens mis en œuvre n'ont cessé d'augmenter.

On peut considérer plusieurs domaines issus ou en relation avec l'e-santé ou la télésanté:

- la télémedecine, qui a pour objectif d'utiliser les moyens de télécommunication pour la réalisation ou l'assistance à l'exercice de la médecine. Elle est considérée par l'OMS comme une composante à part entière de la médecine. Ses champs d'action sont de deux natures, premièrement lié à l'éloignement patient-praticien ou encore liés au diagnostic utilisant les TIC (service de téléradiologie) ;
- l'informatique médicale, qui peut être considérée comme un sous-domaine de la télésanté, vise à utiliser les TIC mais au sens large, pour la surveillance épidémiologique, la promotion de la santé, le traitement et le stockage de l'information médicale. Cette discipline, bien qu'éloignée du patient, a un impact de plus en plus indispensable sur l'organisation générale des soins.

Décret sur la télémedecine

Le gouvernement français, au mois d'octobre 2010 a publié un décret relatif à la télémedecine [6] qui modifie le code de la santé publique avec l'ajout d'un chapitre dédié [7].

Ce chapitre définit les éléments qui peuvent constituer un acte de télémedecine. Ils sont au nombre de 4, avec la téléconsultation, la téléexpertise, la télésurveillance médicale et la téléassistance médicale. La principale évolution concerne évidemment la prise en charge financière d'un acte de télémedecine (sous certaines conditions).

Ce décret va, dans les années à venir, bouleverser l'organisation de la médecine mais de nombreux défis technologiques sont levés par ces dispositions. En effet, les moyens techniques devront permettre la réalisation de ces actes médicaux dans de bonnes conditions de confort et de sécurité.

1.1.4. Dispositions légales pour le partage de données électroniques de santé

Pour des raisons évidentes de sécurité, la télétransmission de données de santé est soumise à un cadre légal strict dans les différentes gouvernances européennes. Cette partie a pour objectif de faire l'inventaire de ces différentes dispositions afin d'effectuer les bons choix applicatifs, techniques, technologiques et architecturaux pour la création d'un réseau d'échange de données médicales.

1.1.4.1. Historique Français

La France s'est souvent reprochée d'avoir, par le biais d'un fichage systématique (Fichier Tulard), largement contribué à faciliter l'organisation de la déportation. Dans le domaine de la vie privée, ces

heures les plus sombres de la société française ont mené à une protection maximale des données et de leur traitement.

1.1.4.2. Informatique et libertés

Le Système automatisé pour les fichiers administratifs et le répertoire des individus (SAFARI), grand projet d'interconnexion de fichiers nominatifs par l'administration française a précipité la création de la « Loi relative à l'informatique, aux fichiers et aux libertés du 6 janvier 1978 ». Cette loi encadre le traitement de l'information en France, en stipulant que l'informatique « *Ne doit porter atteinte ni à l'identité humaine, ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles ou publiques* ». Cette loi instaure aussi des droits aux citoyens, en leur donnant accès (opposition/rectification) aux données les concernant. Elle a aussi placé la France, avec la Suède, parmi les premiers pays à adopter un pareil dispositif.

Parallèlement, la CNIL¹ est créée pour veiller à son application. C'est une entité indépendante qui a deux missions principales :

- assurer le recueil des différentes bases de données à caractère nominatif. La personne responsable de cette base doit en faire la déclaration ;
- veiller à ce que ces bases respectent la loi en vigueur (par contrôles etc.)

La déclaration d'une base de données n'est pas toujours une simple formalité. En fonction du caractère de la base, la CNIL, peut refuser son utilisation, si elle comporte des données de certaines natures. Sont exclues d'un stockage informatisé les données raciales, religieuses, de tendance politique ou encore les informations médicales. Ces dispositions sont censées protéger la confidentialité des personnes et protéger leur vie privée. Néanmoins, l'origine ethnique ou l'orientation religieuse peut être une information extrêmement importante : pour la prise en charge médicale par exemple, la transfusion sanguine est proscrite pour les témoins de Jéhovah, pour l'analyse épidémiologique, l'origine ethnico-raciale peut avoir un impact important.

Dans ces circonstances, la CNIL peut accorder des dérogations mais le dossier est soumis à une demande d'autorisation et non plus une simple déclaration. Il en est de même pour le transfert de données d'identité ou médicales nominatives au travers de réseaux. En effet, pour ce genre de données extrêmement sensibles, la CNIL étudie de façon très poussée les dossiers et statue sur la recevabilité de ceux-ci en fonction du cadre légal. Sans cette bénédiction, il est illégal en France d'effectuer le traitement indiqué.

Dans la majorité des cas, la CNIL ne fait que référencer le contenu des bases de données et ne s'oppose que rarement au traitement de l'information. Dans d'autres cas, lorsque le traitement porte sur des données sensibles, la CNIL n'a pas qu'un simple avis à émettre mais une autorisation, sous forme de numéro d'agrément CNIL de sa part est nécessaire pour procéder au traitement. Si le champ d'application est étendu ou que la nature du traitement change, il est nécessaire de créer un avenant au dossier qui est alors soumis à approbation.

¹ Commission Nationale Informatique et Libertés

Dispositions françaises

Le gouvernement français, avec la loi Informatique et Libertés [1.1.4.2], s'est naturellement protégé contre ces risques potentiels liés à l'échange de données nominatives sur un réseau. Cette loi [8] stipule dans son article 8:

Il est interdit de collecter ou de traiter des données à caractère personnel qui font apparaître, directement ou indirectement, les origines raciales ou ethniques, les opinions politiques, philosophiques ou religieuses ou l'appartenance syndicale des personnes, ou qui sont relatives à la santé ou à la vie sexuelle de celles-ci.

Cette disposition est très claire et interdit le traitement de données entre autres de santé à caractère personnel. Néanmoins, la suite de l'article présente tout un ensemble de conditions selon lesquelles le traitement n'est pas soumis à interdiction parmi celles qui nous intéressent ici :

- *Les traitements pour lesquels la personne concernée a donné son consentement exprès [...]*
- *Les traitements nécessaires aux fins de la médecine préventive, des diagnostics médicaux, de l'administration de soins ou de traitements, ou de la gestion de services de santé et mis en œuvre par un membre d'une profession de santé [...]*
- *Les traitements nécessaires à la recherche dans le domaine de la santé [...]*
- *Si les données à caractère personnel visées sont appelées à faire l'objet à bref délai d'un procédé d'anonymisation [...]*
- *Les traitements, automatisés ou non, justifiés par l'intérêt public [...]*

Ce résumé de dérogations, relatives à la santé, explique les conditions selon lesquelles un traitement de données à caractère médical peut être mis en place, et ceci de manière automatique/informatisée ou non.

Ainsi, la première condition semble être la plus évidente : un traitement sur les données peut être effectué si la personne concernée a donné son consentement explicite. Le problème majeur de cette mesure est bien entendu de contacter, recueillir et obtenir un consentement qui est une opération délicate, chronophage présentant un taux de réponses positives assez faible. En effet, la demande explicite d'un consentement recueillera inévitablement un taux de réponses positives plus faible qu'une clause où le consentement est implicite où la personne possède un droit d'opposition, ce qui est actuellement le cas de l'immense majorité de systèmes d'information médicaux.

D'autres dérogations sont accordées lorsque le traitement des données permet d'assurer la continuité des soins du patient ou encore justifiées d'intérêt public (pandémie...). Cependant, certaines clauses sont intéressantes pour la mise en place d'un réseau : le traitement, s'il est rendu anonyme ou mieux encore, s'il est nécessaire à la recherche dans le domaine de la santé peut s'effectuer. La question est alors : *comment justifier cette nécessité ?*

La justification d'une recherche nécessaire à la santé est soumise, suivant l'article 54, à un comité consultatif mandaté par le ministère de la santé [9], chargé de donner un avis sur la pertinence du volet recherche dans le dépôt de la demande. A défaut de réponse dans les délais d'un mois la réponse est considérée positive.

Les recherches dans le domaine biomédical doivent adhérer aux directives présentes dans le document de référence [10]. Cependant il est possible de prétendre à une procédure « accélérée » permettant de répondre rapidement à des cas d'urgence (épidémiologie).

1.1.4.3. La directive Européenne 95/46/CE

Dans la lignée de la loi Informatique et Libertés, l'Union Européenne s'est dotée, en 1995 d'un texte commun pour l'ensemble des pays membres [11]. Cette loi est largement inspirée du texte français mais est moins restrictive : elle n'impose pas une organisation de contrôle (indépendante) comme la CNIL. Cependant de nombreux pays voisins se sont dotés d'une structure similaire [12]. Les précurseurs de la protection de la vie privée en Europe sont certains *lands* allemands avec des décisions juridiques dès 1970 (Hesse et Bavière). Cependant la première véritable institution nationale est apparue en suède en 1973, puis en France en 1978. Depuis, les autres pays européens ont emboité le pas : *Bundesbeauftragte für den Datenschutz und die Informationsfreiheit* (DE-1978), *Information Commissioner* (UK-1984), *Agencia Española de Protección de Datos* (ES-1993), ou encore *Datenschutzkommission* (AT-1999). L'UE s'est aussi dotée d'une structure similaire, le CEPD¹ qui a aussi un devoir de coopération inter-état. La [Figure 1] représente de façon partielle l'agencement au niveau Européen des instituts de protection des données.

Cette directive a un impact très important sur le développement de l'e-santé au sein de l'UE. Les mesures prises pour gérer les diverses mesures de sécurité, pour s'assurer que la confidentialité du patient est bien assurée ou aussi pour garantir des moyens de traçabilité de l'accès à l'information figurent dans ce texte.

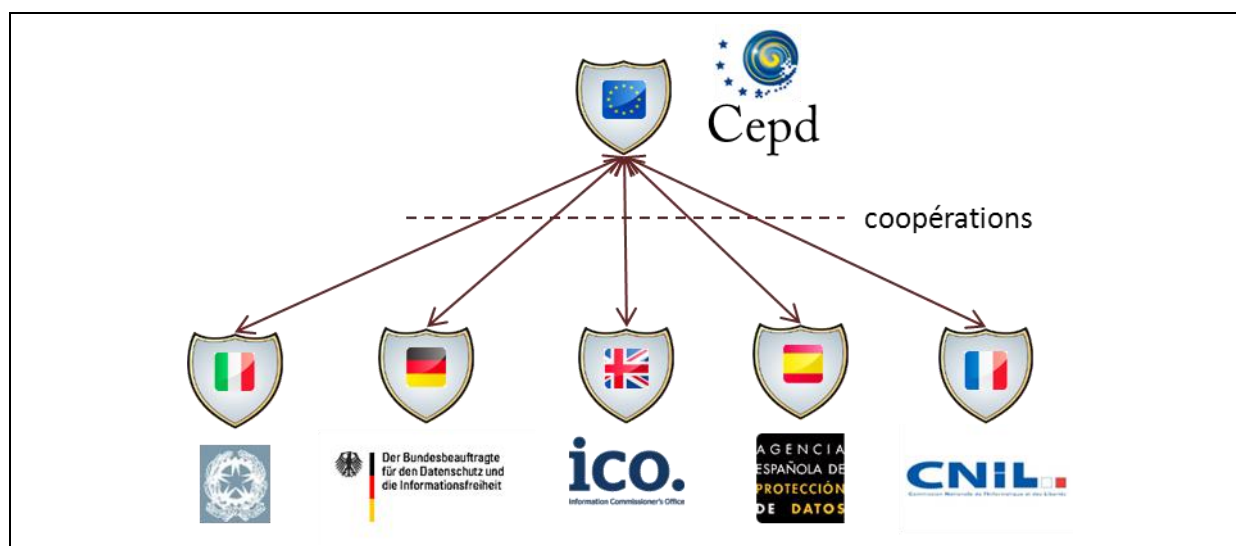


Figure 1 - Organisation Européenne de la protection des données

¹ Contrôleur Européen de la Protection des Données

Aperçu des directives européennes

Suivant les directives européennes, [11] ayant validité, en France, la qualité des données à caractère personnel implique, qu'elles doivent être [13]:

- traitées loyalement et licitement ;
- collectées pour des finalités déterminées, explicites et légitimes, et ne pas être traitées ultérieurement de manière incompatible avec leurs finalités ;
- adéquates, pertinentes et non excessives au regard des finalités pour lesquelles elles sont collectées et pour lesquelles elles sont traitées ultérieurement ;
- exactes et, si nécessaire, mises à jour ;
- conservées sous une forme permettant l'identification des personnes concernées pendant une durée qui n'excède pas celle nécessaire à la réalisation des finalités pour lesquelles elles sont collectées et pour lesquelles elles sont traitées ultérieurement.

Bien entendu, l'obligation de secret doit être respectée quand le traitement porte sur des données à caractère médical, ou a toute autre caractère sensible (racial, ethnique, religieux, ...) de plus, le personnel ayant accès à ces données doit impérativement être sous couvert d'une attestation de respect du secret médical.

Ensuite, un des volets les plus importants est de fournir à la personne concernée :

- un droit d'information sur le traitement effectué ;
- un droit d'accès de modification/rectification sur les données le concernant ;
- un droit d'opposition au traitement.

Concernant la sécurité des données, celle-ci doit être de plusieurs natures :

- protection des données : mettre en œuvre par les moyens techniques appropriés une protection contre la destruction, la perte ;
- sécurité des accès : protéger de l'accès par des personnes non autorisées (très important au travers d'un réseau), et cela en concordance avec le niveau de sensibilité des données concernées ;
- assurer la sécurité contre les intrusions, contre l'accès illicite aux données, protéger les échanges sur le réseau par les outils adéquats.

1.2. ORGANISATION DU DEPISTAGE ORGANISE DES CANCERS EN FRANCE

1.2.1. *Le cancer*

1.2.1.1. *La maladie*

Le cancer est une pathologie génétique liée à la dégénérescence cellulaire d'un individu. C'est plus précisément le terme générique utilisé pour décrire le phénomène où des cellules de l'organisme se comportent de manière anormale à cause d'une modification de leur génome et d'accumulation au niveau de l'ADN des mutations nécessaires à la multiplication cellulaire. C'est une

maladie à évolution le plus souvent lente, qui se concentre au début autour d'un tissu de l'organisme. Dans des phases plus avancées de la maladie, les cellules cancéreuses peuvent migrer à d'autres tissus, organes, jusqu'à l'ensemble du corps de la personne par le phénomène appelé métastase. On parle dans ce cas de cancer généralisé.

Globalement, les chances de guérison sont inversement proportionnelles à la précocité de la détection de la maladie.

1.2.1.2. *L'anatomo-pathologie*

L'anatomo-pathologie, la cytopathologie ou l'anatomo-cyto-pathologie est une discipline médicale spécialisée dans l'analyse in vitro de tissus organiques ou pièces opératoires afin d'en déceler le caractère pathologique ou non. Les laboratoires d'anatomie-pathologie forment une pièce maîtresse dans le dépistage du cancer puisqu'ils assurent la quasi-exhaustivité des diagnostics. Seuls sont exempts de diagnostic pathologique les maladies sanguines (leucémies) ou d'origine génétique.

Un compte-rendu pathologique présente de manière textuelle le résultat de l'analyse en laboratoire de la pièce réceptionnée. Il s'en suit une (ou plusieurs) codifications de la lésion observée (en cas de pathologie avérée) au format ADICAP.

L'ADICAP

La codification ADICAP, qui porte de nom de l'association qui promeut ce format [14] est une classification normalisée des lésions cancéreuses. Elle est diffusée sous licence « *Creative Commons* » et nécessite une simple adhésion pour pouvoir l'utiliser. Cette codification alphanumérique permet de représenter simplement l'ensemble de la chaîne de traitement pathologique d'une pièce opératoire. Elle permet de savoir le mode de prélèvement (ponction, aspiration, curetage, ...), la technique de diagnostic utilisée (histologie, imagerie, immuno-histochimie, ...), l'organe concerné par l'examen, puis un ensemble de sigles représentant la pathologie générale, tumorale et sa gravité.

Ainsi, le code **FIGX0I80** désigne l'analyse par immuno-histochimie(I) d'un frottis(F) issue de la région cervico-vaginale(GX) qui présente une lésion communiquée (OI80). Un autre exemple **OHDC A7A0** est une analyse par histologie (H) de l'exérèse(O) d'un colon(DC) qui présente un adénocarcinome invasif. En d'autres termes un cancer avéré du colon.

Cette codification, malgré son apparente simplicité, permet de fournir un ensemble d'informations essentielles à l'étude épidémiologique de la maladie. De plus, le format est reconnu comme un standard national de la description des lésions cancéreuses puisqu'il figure sur la quasi-totalité des rapports anatomo-pathologiques. Par ailleurs, cette donnée est le plus souvent structurée au sein des systèmes d'information des laboratoires contrairement au compte rendu médical textuel qui est difficilement exploitable. De plus, en cas d'absence de cette codification, tout pathologiste pourra de façon assez précise retrouver ce code à partir du diagnostic textuel.

Notons que cette codification est obligatoire sur les 8 premiers caractères mais a été étendue à 15 afin de permettre d'informer plus précisément sur le diagnostic pathologique comme par exemple la codification de lésions apparues depuis une autre origine (extension hépatique d'une lésion à la vésicule biliaire par ex).

De plus, la codification Adicap est interchangeable facilement, et avec un taux de perte d'information faible avec la codification OMS des pathologies [15].

1.2.2. Les sources de données sur le cancer

Du fait de l'étendue de la maladie, des moyens de traitement comme de prise en charge, les sources de données sont nombreuses et variées. Néanmoins on peut distinguer plusieurs types de sources de données : les sources médicales en relation directe avec le patient et l'établissement de soin et des sources plus globales servant aux études statistiques. Bien entendu ces dernières sont le plus souvent alimentées par les premières.

1.2.2.1. Les sources de données médicales

L'anatomie pathologique

En premier lieu, comme évoqué précédemment, on citera les laboratoires et services de pathologie. La quasi-exhaustivité des cancers a un compte rendu pathologique mais il n'existe aucun moyen généralisé de recueil de ces informations, qui restent le plus souvent peu exploitées à grande échelle.

Les données PMSI

Le Programme de Médicalisation des Systèmes d'Information a permis d'instaurer, au sein de tout établissement de santé, l'élaboration d'une fiche de résumé sur tout passage d'un patient dans le système de soin. C'est une base centralisée par la DHOS¹ directement au ministère de la santé. Cependant l'accès à ces données n'est pas très aisé et beaucoup de travail de retraitement est nécessaire pour pouvoir l'exploiter efficacement.

Les données ALD

L'Affection Longue Durée est une reconnaissance par l'assurance maladie d'une pathologie chronique. De ce fait, les personnes répertoriées ALD bénéficient d'une prise en charge de l'intégralité de leurs soins par celle-ci. La base ALD30 est la liste des pathologies concernées qui comportent un traitement thérapeutique prolongé ou un coût de prise en charge élevé. La plupart des cancers y figurent, aux côtés de pathologies telles que la maladie d'Alzheimer ou Parkinson, le diabète ou les insuffisances respiratoires. Cette liste n'est en aucun cas exhaustive et d'autres pathologies, suivant leur forme et leur caractère peuvent être considérées ALD.

L'assurance maladie

La France dispose depuis 1945 de l'assurance maladie qui prend en charge une partie des frais de santé de ses concitoyens. Le recul que présente cette source de données sur l'état de santé du pays est unique au monde. Malheureusement l'accès aux données des diverses caisses d'assurance maladie n'est pas généralisé et ne se fait qu'au compte-goutte, le plus souvent uniquement par les services de l'Etat (INSEE) ou par certains instituts de recherches INSERM spécialisés en épidémiologie. Cependant ces organismes sont plus habitués à fonctionner par étude de type cohorte.

¹ Direction de l'Hospitalisation et de l'Organisation des Soins

1.2.2.2. *Les autres sources de données*

Le Centre de Regroupement Informatique et Statistique en Anatomie et cytologie Pathologiques (CRISAP)

Les CRISAPs sont des structures médicales regroupant les anatomo-pathologistes d'une même région afin de collecter les données à des fins d'exploitation statistique par les pathologistes eux-mêmes. Les CRISAPs disposent d'une fédération nationale regroupant les différentes initiatives régionales. Malheureusement ce système avait un défaut majeur qui imposait au médecin pathologiste de d'envoyer manuellement ses données au serveur (déclaration) avec toutes les contraintes que cela comportait (erreurs humaines, temps médecin alloué, perte de contrôle sur les données). Ainsi, au vu du nombre de participants se restreignant, les CRISAPs sont peu à peu tombés en désuétude et seul un faible nombre sont encore en activité. De plus, la qualité des données n'était pas optimale car l'envoi est effectué de façon anonyme : aucun nettoyage de doublons n'est effectué.

Un Crisap existe en Auvergne, hébergé par le CHU de Clermont-Ferrand. Cependant, il n'est plus en activité par manque de participation des pathologistes.

Les registres du cancer

Un des moyens les plus efficaces est de répertorier de façon systématique tous les cancers et d'assurer leur suivi. Ainsi, certaines régions françaises se sont dotées d'un tel système, en collaboration avec l'ensemble des professionnels de santé concernés. Ces registres donnent un bon aperçu de l'activité locale mais les efforts humains et financiers à déployer pour collecter, nettoyer et analyser toutes ces données empêchent la généralisation de ce type de système. Le principal avantage d'un registre est de collecter, de manière systématique les incidences et mortalités des cancers.

En France [16], environ 20% de la population est couverte par un registre du cancer qui sont au nombre de 26 qualifiés par le CNR¹ en France métropolitaine auxquels s'ajoutent un registre qualifié en Martinique, un registre spécialisé en Guadeloupe et un registre en cours de certification en Guyane Française, répertoriés en [Figure 2].

¹ Centre Nationaux de Référence

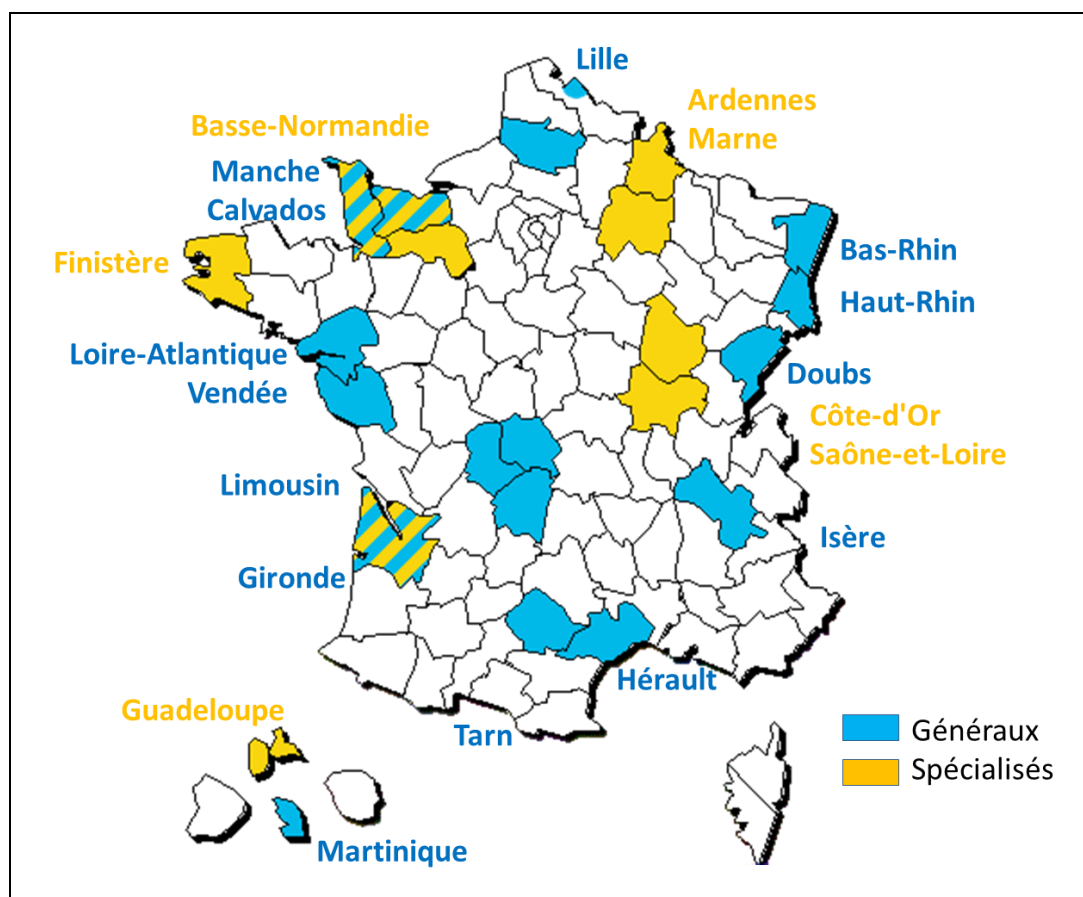


Figure 2 - Les registres des cancers en France. Source Francim¹

Les données de mortalité

L'examen des actes de décès et de la cause de mortalité est un indicateur statistique assez fiable sur l'ensemble du territoire. La standardisation CIM-10 [17] permet de coder les causes médicales de décès. Le CépiDc [18] est une structure de l'INSERM qui produit tout un ensemble de statistiques épidémiologiques en collectant les données sur les sources de décès en France.

1.2.3. L'épidémiologie et la veille sanitaire

L'étude épidémiologique du cancer est une nécessité primordiale pour combattre la maladie. Le cancer est ainsi devenu, en France, en 2007 la première cause de mortalité chez les hommes et la seconde chez les femmes derrière les maladies de l'appareil circulatoire (cardiopathies, hypertension, rhumatismes). La veille sanitaire influence de façon directe la politique de santé, comme l'atteste les différents « plans cancer » 2003-2007 et 2009-2013.

Cependant, l'étude épidémiologique du cancer se heurte à de nombreux écueils, notamment en France sur la collecte des données. Bien que le système de soin français soit relativement moderne, les difficultés de collecte de données médicales sont nombreuses. Ainsi, les sources mises à disposition des épidémiologistes ne sont pas homogènes sur le territoire et rarement exhaustives. Les registres des cancers, censés répertorier l'intégralité des tumeurs dans une région spécifique ne

¹ Réseau de regroupement des registres des cancers

couvrent qu'une partie de la population. Les autres sources, comme ALD30¹ ne comprennent que les personnes qui se sont volontairement inscrites, les données PMSI² sont trop disparates pour avoir une situation géographiquement fiable et les données de l'assurance maladie ne sont que rarement divulguées aux épidémiologistes.

Les seules données de 'bonne' qualité sur l'ensemble du territoire sont les informations de mortalité, issues des actes de décès des personnes. Mais qui ne permettent évidemment pas de connaître ni l'incidence, ni la prévalence des cancers.

1.2.4. Le dépistage

Le dépistage du cancer a plusieurs objectifs [19], il s'agit, pour des pathologies dites curables de les détecter de façon spontanée et le plus précocement possible afin d'augmenter significativement les chances de guérison.

1.2.4.1. Moyens du dépistage

Il existe différentes manières de dépister le cancer :

- l'examen physique : il s'agit de détecter visuellement les signes pathologiques (grosseurs, hématomes) avec l'appui de l'historique médical du patient ;
- les tests en laboratoire : partant d'une pièce opératoire à risque (tissu, sang, prélèvement d'organe) il s'agit d'effectuer une analyse histologique et d'en déduire un diagnostic pathologique ;
- l'imagerie médicale : permet de détecter à l'intérieur du corps du patient des anomalies. Cette étape n'est souvent qu'une étape préliminaire à une analyse pathologique ;
- la génétique : certaines mutations génétiques sont clairement liées à certains types de cancers.

Le dépistage a été reconnu comme solution efficace pour réduire de manière significative la mortalité de certains cancers dits « évitables ».

1.2.4.2. Cancers concernés par le dépistage organisé

Origine

Au sein de l'UE [20], le constat est assez simple, on voit, dans les [Figure 3] et [Figure 4] que les cancers du sein, colon et col utérin représentent chez la femme presque la moitié des cas de cancer pour seulement un tiers des décès. Il en est de même pour le cancer du colon chez l'homme, avec une mortalité proportionnellement inférieure à l'incidence.

Ces cancers sont considérés comme 'évitable' car les chances de guérison en cas de dépistage précoce sont élevées.

¹ Affection Longue Durée

² Programme de Médicalisation des Systèmes d'Information

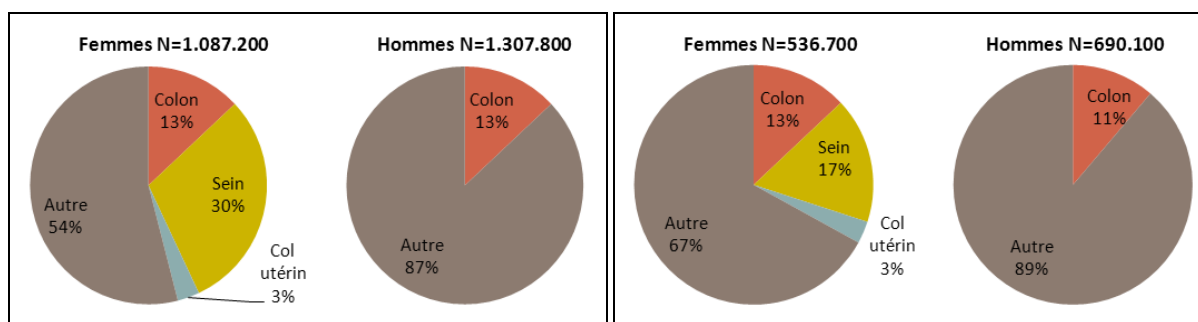


Figure 3 - Cas de cancer du sein, colon et col utérin en UE (2007) [20]

Figure 4 - Décès dus aux cancers du sein, colon et col en UE (2007) [20]

1.2.4.3. Organisation du dépistage organisé

Fonctionnement

Le dépistage organisé est la généralisation à une catégorie de la population d'une zone géographique définie du dépistage d'une maladie. Elle se déroule le plus souvent par sensibilisation et invitation des personnes concernées. La généralisation d'une procédure de dépistage permet de baisser significativement la mortalité de ces cancers.

A l'heure actuelle, le cancer du sein et du colon a été repris dans les programmes de dépistage organisé dont le cahier des charges fait l'objet d'un arrêté ministériel [21]. Les programmes sont régis au niveau national par le ministère de la santé qui délègue, en région, aux ARS leur déploiement. Celles-ci mandatent des associations départementales créées pour assurer la mise en œuvre des programmes.

La mise en œuvre du dépistage organisé est donc confiée à ces associations qui ont alors la charge d'inviter la population ciblée à se faire dépister à intervalle régulier et aussi (et surtout) d'assurer le suivi des personnes dépistées en collectant les données médicales issus des examens passés suite au dépistage (positif ou négatif).

Historiquement, le cancer du sein a été le premier dépisté, est venu ensuite le cancer colorectal puis celui du col utérin qui est en expérimentation depuis 2010 (mais pas encore généralisé). D'autres types de cancers sont concernés par le dépistage, mais celui-ci n'est ni organisé ni généralisé, on peut citer ceux liés à la peau (mélanome) ou à la cavité buccale.

1.2.4.4. Estimation quantitative des cancers dépistés

La [Figure 5], présente une estimation quantitative des différents cancers concernés par le dépistage organisé. Là où l'estimation des populations est assez aisée, puisqu'elle repose sur un recensement national de la population, l'estimation de l'incidence et de la mortalité des cancers est beaucoup moins évidente. Les données présentées pour l'Auvergne sont à prendre avec beaucoup de précaution. Les raisons sont dues à un manque de registre du cancer local, et à une faible statistique locale. En effet, seuls 2270 cancers du sein ont été recensés sur le territoire en 2007 pour moins de 300 décès.

Concernant le cancer colorectal et du col utérin les données à dispositions ne permettent pas d'avoir une estimation fiable des taux de réponses au dépistage. En effet, la généralisation de ces campagnes n'a été faite qu'en 2009. Cependant, on peut estimer suivant l'INCA [16] que le taux de

réponse aux campagnes de dépistage du cancer colorectal pour le département de l'Allier est supérieur à 50% et situé entre 40 et 45% pour le Puy-de-Dôme lors de la première campagne de 2007.

Les informations disponibles proviennent des structures de dépistage qui produisent chaque année un rapport d'activité, comprenant les dépistages effectués, les relectures ou seconds diagnostics ainsi que les cas positifs détectés. D'autres sources, comme les données PMSI ou d'assurance maladie permettent d'obtenir une estimation. Cependant, compte tenu de la faible participation au dépistage, avec une moyenne autour de 50% sur le territoire en 2007 [22], du manque de données sur les personnes déjà traitées pour un cancer (ALD) conjugué au manque d'informations fiables sur les cancers de l'intervalle (entre deux dépistages) cette estimation ne peut être qu'approximative.

Le rôle épidémiologique du dépistage organisé est pourtant reconnu et a son importance. C'est en effet le seul programme à l'échelle nationale qui permet de connaître, même en partie, l'état de santé de la population vis-à-vis des cancers dépistés.



Type de Cancer	Population ciblée	Estimation de la population ciblée (2007) ¹			Incidence Mortalité	
			En France	En Auvergne	En France (2005) [23]	En Auvergne (2007) [16] ²
Sein 	Toutes les femmes	Population	9,2 M	206.000		
	âgées de 50 à 74 ans	Invitations	4,3 M	102.000	49.814	2270
		Répondants	2,18 M	56.000	11.201	278
Colorectal 	Toutes les personnes âgées de 50 à 74 ans		17,8M	389.000	37.413	3878
					16.865	451
Col utérin	Toutes les femmes		17,6M	420.000	5.774	878
	entre 25 et 65 ans				1.800	11

Figure 5 - Estimation de la population française ciblée par le dépistage organisé des cancers

La difficulté d'estimation de l'incidence de la maladie est bien réelle, en particulier pour la région Auvergne. Seule une petite partie de la population française est couverte par un registre des cancers. Les études épidémiologiques doivent alors se fonder sur des estimations et extrapolations statistiques issues des données des registres [24], ou sur la base ALD30/PMSI.

1.2.4.5. Acteurs du dépistage organisé

En France, le dépistage est réalisé par le biais d'un système associatif, mandaté par l'Etat en collaboration avec les caisses d'assurance maladie et les différents corps médicaux. Ces associations ont une mission complexe : elles doivent inviter une population à se faire dépister, effectuer une relecture des diagnostics (mammographies), assurer le suivi des patients, publier des statistiques sur l'efficacité du dépistage : taux de réponse, cancers diagnostiqués. Le rôle épidémiologique est aussi

¹ Source : INSEE ; Recensement de la population 2007

² Source : Cépidic ; données 2007

extrêmement présent puisqu'elles peuvent répondre de façon précise et rapide aux taux d'incidence et de prévalence de ces cancers.

Ces actions nécessitent de nombreux moyens humains et techniques afin d'assurer la collecte des informations nécessaires. La qualité et l'exhaustivité du travail influencent fortement les « performances » de l'association, qui sont, à ce propos, reprises lors de leur évaluation annuelle.

1.2.4.6. Fonctionnement du dépistage organisé

La [Figure 6] représente sous forme d'un diagramme de séquence le fonctionnement des associations de dépistage organisé des cancers pour le sein. Le fonctionnement est similaire (sauf pour le mode de dépistage) pour tout type de cancer.

La première étape consiste à envoyer, à chaque femme concernée par le dépistage (données fournies par l'assurance maladie), une invitation papier au dépistage. Cette invitation contient une liste de cabinets de radiologie agréés pour effectuer la mammographie dont la prise en charge par la sécurité sociale est complète. Cette mammographie est par la suite analysée par un radiologue qui définit son caractère pathologique (ou non). Une deuxième lecture de cette imagerie est effectuée par un autre radiologue rattaché à la structure de dépistage. Si le résultat est positif, la patiente est convoquée à un examen pathologique approfondi, qui consiste en une exérèse de la zone concernée puis analyse en laboratoire pathologique. Si le résultat s'avère positif la patiente est prise en charge, dans le cas contraire la patiente reste dans le circuit normal du dépistage.

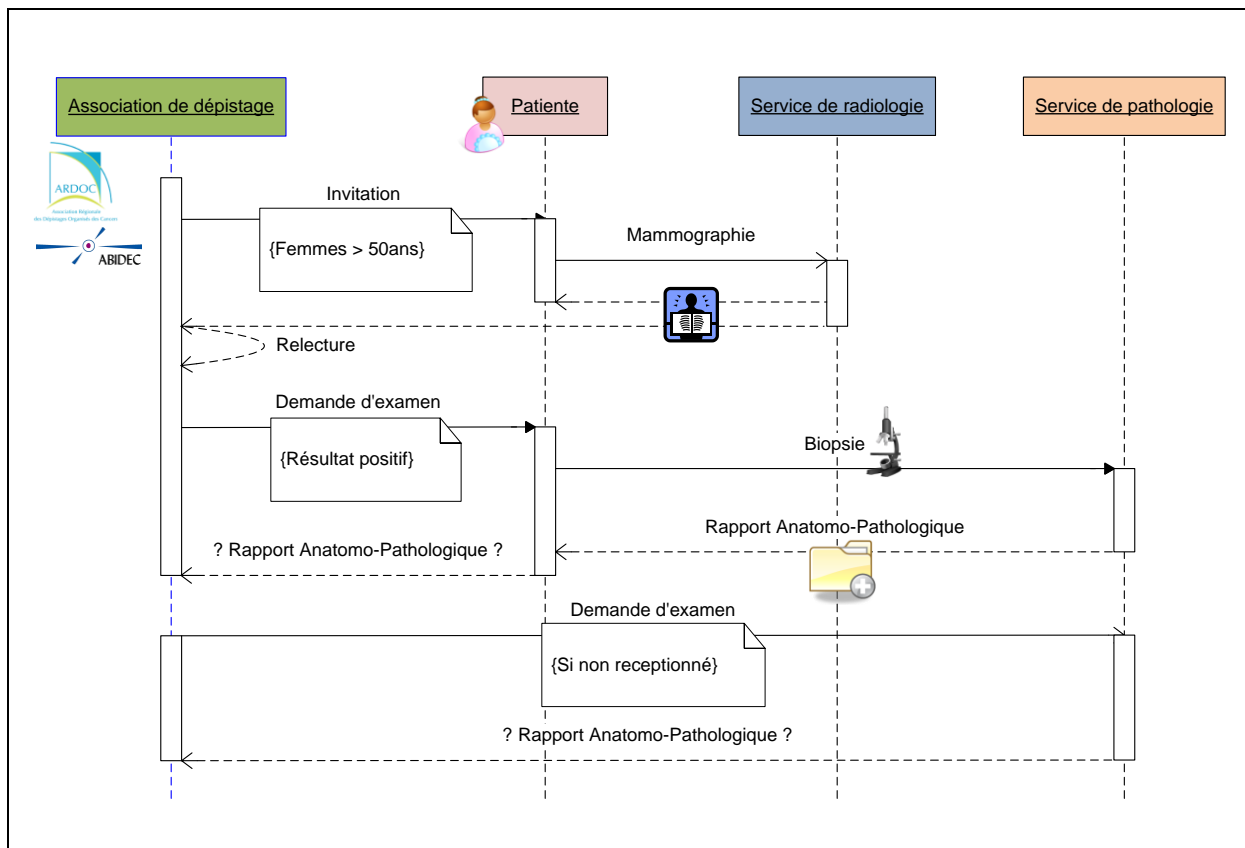


Figure 6 - Diagramme de séquence du dépistage organisé du cancer du sein

Le principal défaut de cette séquence se situe au niveau du retour à l'association de dépistage du rapport anatomo-pathologique. En effet, seule la patiente reçoit de manière systématique le rapport pour le transmettre à son médecin prescripteur. L'association doit alors demander soit à cette

dernière soit directement au service de pathologie à récupérer ce rapport – indispensable pour assurer le suivi des patientes dépistées.

1.2.5. Efficacité du dépistage organisé

L'efficacité du dépistage organisé est encore difficile à évaluer. L'efficacité de dépistage du cancer du col utérin est très disparate dans les pays de l'UE ayant ou non inscrit le dépistage organisé dans leur politique de santé publique [25]. C'est tout le dispositif qui est parfois remis en question, par exemple, l'impact du dépistage organisé du cancer du sein sur la mortalité n'a pas été significatif au Danemark [26] mais le contraire a pourtant été démontré en Suède [27].

De plus, la comparaison entre les efforts faits par la France, où le nombre d'appareils de mammographie est supérieur à 80/million de femmes [28] (ce qui en fait le 3^{ème} pays au monde, derrière l'Autriche et les USA), ne montre pas de baisse spectaculaire de la mortalité comparé au Royaume-Uni [29], qui ne dispose d'un taux d'équipement que de 21/million sur la même période.

Malgré toutes ces controverses, l'efficacité du dépistage semble être acquise et les efforts gouvernementaux des pays occidentaux en la matière peuvent l'attester. Les études épidémiologiques montrent que la qualité des données recueillies peut être aléatoire, en particulier pour la France, où la qualité des données est considérée moyenne [30].

1.3. APPLICATION DES GRILLES POUR LA SANTE

Devant les nombreuses difficultés, que ce soit en termes de qualité des données ou de leur accès- en Auvergne comme ailleurs - le principal objectif de la thèse était de fournir une architecture originale de gestion de données médicales sur la région. En effet, le manque de coordination des sources de données sur le cancer complique fortement le travail des associations de dépistage organisé. De plus, l'absence de structuration et de qualité des données médicales ne permet pas d'effectuer des enquêtes épidémiologiques d'envergure avec une fiabilité suffisante.

1.3.1. Evolution de la gestion de la donnée médicale

1.3.1.1. Historique

Les environnements de gestion de bases de données sont apparus naturellement avec l'informatique. Les bases de l'algèbre relationnelle, posées par EC.Codd en 1970 [31] sont à l'origine des bases de données relationnelles. Elles définissent un ensemble de règles mathématiques entre les relations dans une base de données. La création par la suite des SGBD (Système de Gestion de Bases de Données) a permis techniquement l'implémentation de ces notions de base nécessaires à l'utilisation des bases de données : l'ajout, la suppression, la mise à jour, la recherche, la gestion d'index et tout un ensemble d'outils de manipulation de données sur ces bases.

L'évolution de ces systèmes a donné naissance plus tard, vers la fin des années 80, au concept d'entrepôt de données [32]. Ces systèmes agissent comme un concentrateur de données qui, à

intervalle régulier, collectent un ensemble d'informations depuis des sources. Ces sources sont pour le plus souvent elles-mêmes des bases de données. La plupart du temps, les entrepôts ont une vocation de statistique ou d'analyse. L'avènement des réseaux informatiques et notamment d'internet dans les années 90 a, par la suite, offert de nouvelles possibilités d'échange d'informations, facilitant ainsi la collecte et le transport de données.

1.3.1.2. Origine de la répartition de l'informatique

Face à des problèmes grandissants de manque de puissance des microprocesseurs, le concept d'informatique répartie a fait son apparition. Là où un processeur n'est capable de traiter qu'un certain nombre d'informations à la seconde, un groupe, ou une grappe de n ordinateurs est capable de traiter n fois plus d'informations.

Premièrement dédiée au calcul scientifique (calcul distribué et clusters), la répartition s'est progressivement étendue aux données (datacenters).

Le principal problème de ces centres de calcul et de données est qu'il limite les possibilités de concentration des ordinateurs. Au delà de certaines limites physiques (espace, consommation électrique, miniaturisation et climatisation) ou de maintenance/évolutivité, la concentration n'est plus économiquement viable. C'est alors qu'est né le concept de grille informatique, qui a pour objectif de se servir des réseaux à haut débit pour mutualiser les ressources informatiques existantes. Les capacités de calcul et de stockage offertes par une infrastructure de grille offrent alors une nouvelle dimension aux applications scientifiques, qui sont de plus en plus coûteuses en termes de temps de calcul et en volume de données générées.

La fédération de ces ressources informatiques permet aussi de faire émerger des organisations virtuelles, rassemblées par une thématique scientifique commune pour exploiter ces ressources. La maintenance et la gestion de ce genre d'infrastructure est alors supporté par l'ensemble de la communauté ce qui réduit significativement le coût global d'exploitation par rapport à une solution complètement centralisée.

1.3.2. Les grilles informatiques

Avant-propos

Le terme « grille » a souvent été critiqué comme une traduction littérale du mot « grid » anglophone. On reproche souvent au mot français un manque de profondeur dans sa signification, le terme désignant plus directement une séparation, une barrière ou alors une grille de mots croisés, de sudoku ou de barbecue. Le terme « grid » quant à lui désigne également le réseau de distribution électrique national britannique mais aussi une interconnexion de lignes dans un schéma, en électronique ou encore sur une carte. Ainsi, le terme « computing grid » prend une autre dimension en anglais qu'une « grille informatique ». Néanmoins, le terme français, même sujet à controverse semble être maintenant accepté comme tel, mais toujours accompagné de son qualificatif « informatique » ou, « de calcul ».

1.3.2.1. Historique

Les grilles informatiques ont émergé avec la généralisation des réseaux haut-débit au début des années 2000 pour partager un ensemble de ressources informatiques. Cependant des bases ont été posées dès les années 1960 par F.Corbato [33] à une époque où les ordinateurs avaient un coût prohibitif, d'où le besoin de partage :

« *The time-sharing computer system can unique a group of investigators... one can conceive of such a facility as an... intellectual public utility.* »

ou encore L.Kleinrock, à propos des réseaux, lors de la création d'ARPANET en 1969 [34]:

« *We will perhaps see the spread of 'computer utilities', which, like present electric and telephone utilities, will service individual homes and offices across the country.* »

Ces informaticiens avaient prédit, avant même l'existence des ordinateurs personnels ou des réseaux, l'évolution de l'informatique comme un service global au bénéfice de la société. Par la suite, les réseaux et l'informatique personnelle sont nés, les besoins ont évolué et l'avènement d'Internet a changé la donne. En effet, à la fin du XXème siècle, le constat était simple : d'après les lois de Moore, Gilder et Metcalfe [Figure 7], l'évolution des capacités de calcul vis à vis de l'évolution de celle des réseaux et des capacités de stockage amènerait inévitablement à une décentralisation de l'informatique. En effet, là où Moore, en 1965 [35] affirmait (bien que de façon empirique) que la puissance des microprocesseurs (en terme de nombre de transistors) doublerait tous les 18 mois, Gilder en 1990, quant à lui annonce que les capacités de stockage tripleraient tous les ans tandis que Metcalfe en 1985 a montré que les capacités des réseaux évolueraient de façon proportionnelle à la racine du nombre de nœuds de ce réseau. Lorsque l'on mesure l'expansion que connaît Internet à l'heure actuelle cette définition prend toute sa dimension.

Cependant, de récentes études [36] ont montré qu'à causes de limites physiques de miniaturisation des processeurs, ces lois n'auraient plus cours dans quelques années.

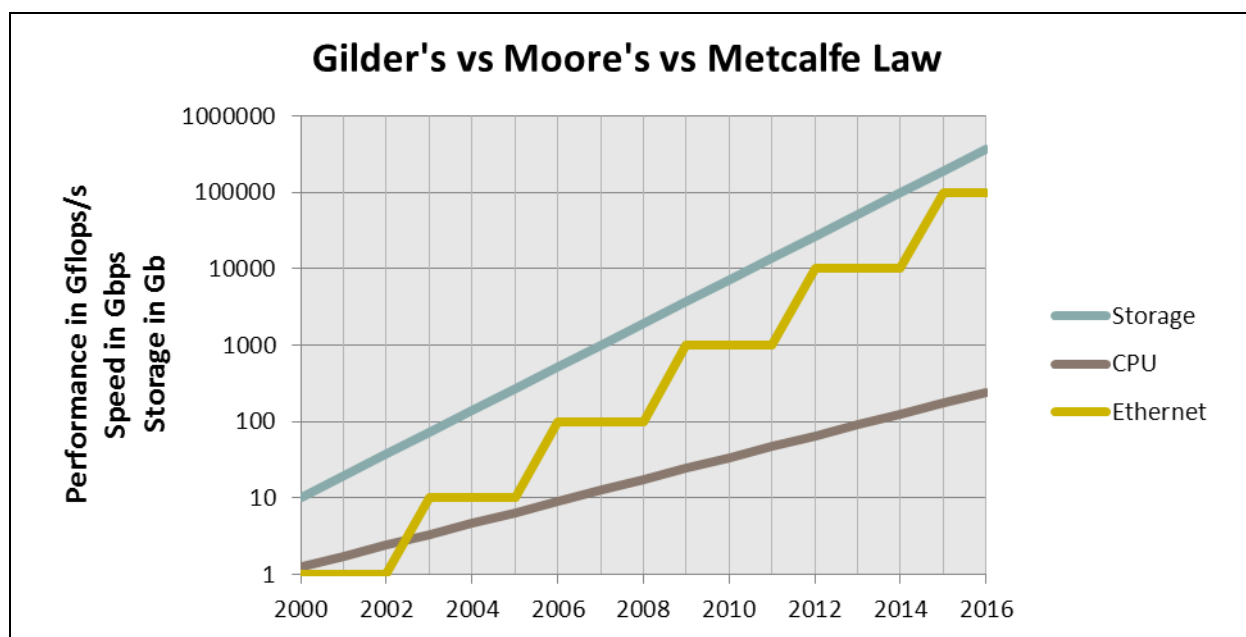


Figure 7 - Comparaison des lois de Moore, Gilder et Metcalfe

L'idée principale d'une grille est d'utiliser un ensemble de ressources informatiques géographiquement dispersées et homogènes à des fins de mutualisation de ces ressources. La définition du grid computing proposée par I.Foster en 1999 [37], considéré comme l'un des principaux acteurs du domaine est :

« *Grid computing is the combination of computer resources from multiple administrative domains for a common goal.* »

Bien que cette définition ait longtemps été sujette à controverse [38], I.Foster proposa de nouveaux concepts en 2002 [39, 40].

1.3.2.2. But des grilles informatiques

En d'autres termes, il s'agit de regrouper au moyen d'une couche logicielle commune, un ensemble de machines reliées entre elles par internet. La principale dichotomie avec le « power grid » anglo-saxon, où l'offre d'électricité est mutualisée sur tout un territoire dans une relation stricte producteur-consommateur; ici, les clients doivent le plus souvent participer (mettre à disposition des ressources) pour pouvoir accéder à celles des autres dans un esprit collaboratif.

Cette principale contrainte est primordiale pour la pérennité d'un tel outil. En effet, si la participation des utilisateurs d'une grille est fortement déséquilibrée, des conflits peuvent alors remettre en question son intégrité. Pour cela, tout un ensemble de règles d'utilisation, de partage ou même de priorité doivent être définies pour maintenir harmonie et équilibre. Ces notions seront développées plus tard dans ce document.

1.3.2.3. Premières implémentations : le CERN

Dès 2001, le Conseil Européen pour la Recherche Nucléaire¹ [41], a eu besoin, lors du lancement du projet LHC (Large Hadron Collider) [42], de trouver un moyen d'exploiter les données qui seront produites par cet accélérateur. En effet, les prévisions du CERN tablaient sur une production de données estimées à 15 Pétaoctets par an. Devant le coût pharaonique d'une infrastructure centralisée capable de stocker puis de traiter cette masse de données, le projet LCG² [43] a été lancé par le CERN.

Partant de l'expérimentation du projet EDG (Datagrid) [44, 45], véritable laboratoire de test d'implémentation de l'infrastructure de grille, LCG a su créer, via les successifs projets EGEE [46] puis EGI [47], un système distribué de gestion de ces données. Le lancement du LHC, dont les premières collisions ont été effectuées le 9 novembre 2009, renforce l'importance de l'infrastructure de grille.

Le rôle du CERN dans le développement des grilles et des intergiciels a été fondamental pour l'avènement de la technologie.

1.3.2.4. Architecture d'une grille informatique

La [Figure 8] représente de façon stratifiée une architecture classique de grille informatique. On peut distinguer quatre couches principales, premièrement le réseau, qui assure les interconnexions des différents nœuds de grilles, représentés sur la deuxième couche. On peut considérer plusieurs

¹ Conseil a été rebaptisé Organisation depuis mais l'acronyme est resté

² LHC Computing Grid

types de ressources, de type 'computing', de stockage ou de capteurs. L'exploitation sous forme applicative de ces ressources par l'utilisateur, en quatrième couche, est gérée par une troisième couche, nommée middleware(intergiciel), qui a pour mission de fournir l'ensemble des outils nécessaires à l'utilisateur pour exploiter les ressources en assurant aussi la sécurité.

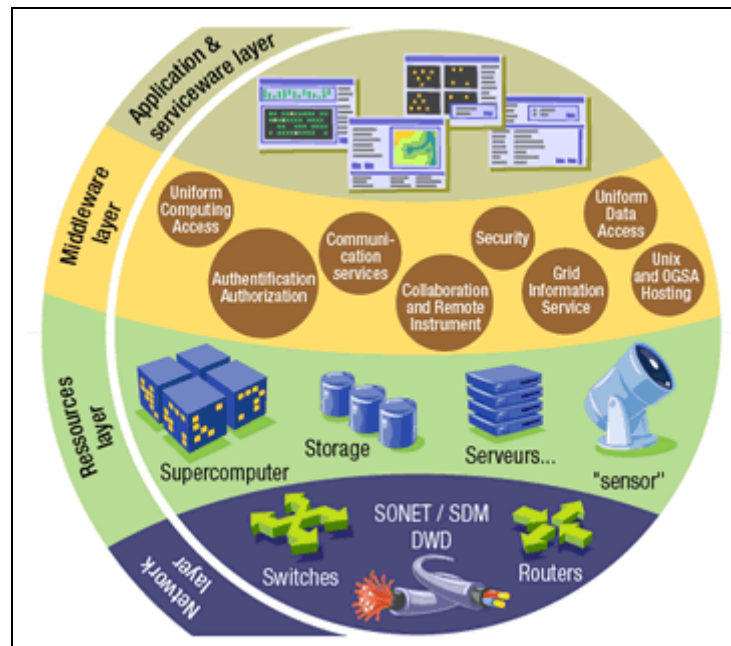


Figure 8 - Architecture d'une grille

1.3.3. Applications des grilles

Le principe même d'une grille est alors de regrouper au sein d'une même entité virtuelle et au moyen d'une couche logicielle commune, un ensemble de ressources informatiques géographiquement distribuées. Les utilisations potentielles de cette entité, appelée « grille informatique » sont alors multiples, du calcul distribué à l'interconnexion de sites en passant par le stockage et le partage de l'information.

1.3.3.1. Interconnexion de sites et partage d'informations – « knowledge grids »

Origine

A l'origine des knowledge grids on peut citer les grilles dites « d'information » qui permettent d'échanger de l'information au travers des réseaux. C'est bien entendu ce concept de grille d'information qui est développé dans Internet, avec une interconnexion mondiale de sites disposant de l'information.

Extension

D'autres applications des grilles sont nées de la prise de conscience des performances des réseaux qui reliaient les différents nœuds informatiques. En effet, ces réseaux dédiés, très haut débit et parfaitement sécurisés par les outils d'authentification, de chiffrement et d'autorisation proposés par les intergiciels de grille ont ouvert de nouvelles perspectives.

En premier lieu, le réseau Européen GÉANT [48], qui vient de fêter ses dix ans, fournit un réseau d'une longueur totale de 50.000km, regroupant 40millions d'utilisateurs issus des infrastructures de recherche de plus de 40 pays.

Le réseau Renater [49] en France, relié au réseau Européen GÉANT sont les supports de ces infrastructures de recherche. Les [Figure 9], [Figure 10] et [Figure 11] montrent l'étendue de ces réseaux et des différents débits, pour Renater tous les nœuds sont connectés en 10Gbit/s.

Au niveau de l'Auvergne, reliée à 10Gbit/s au réseau national Renater, le réseau AuverData alimente les différents sites universitaires régionaux en offrant une bande passante conséquente qui lui permet d'héberger la grille régionale Auvergrid [67] répartie entre plusieurs sites.

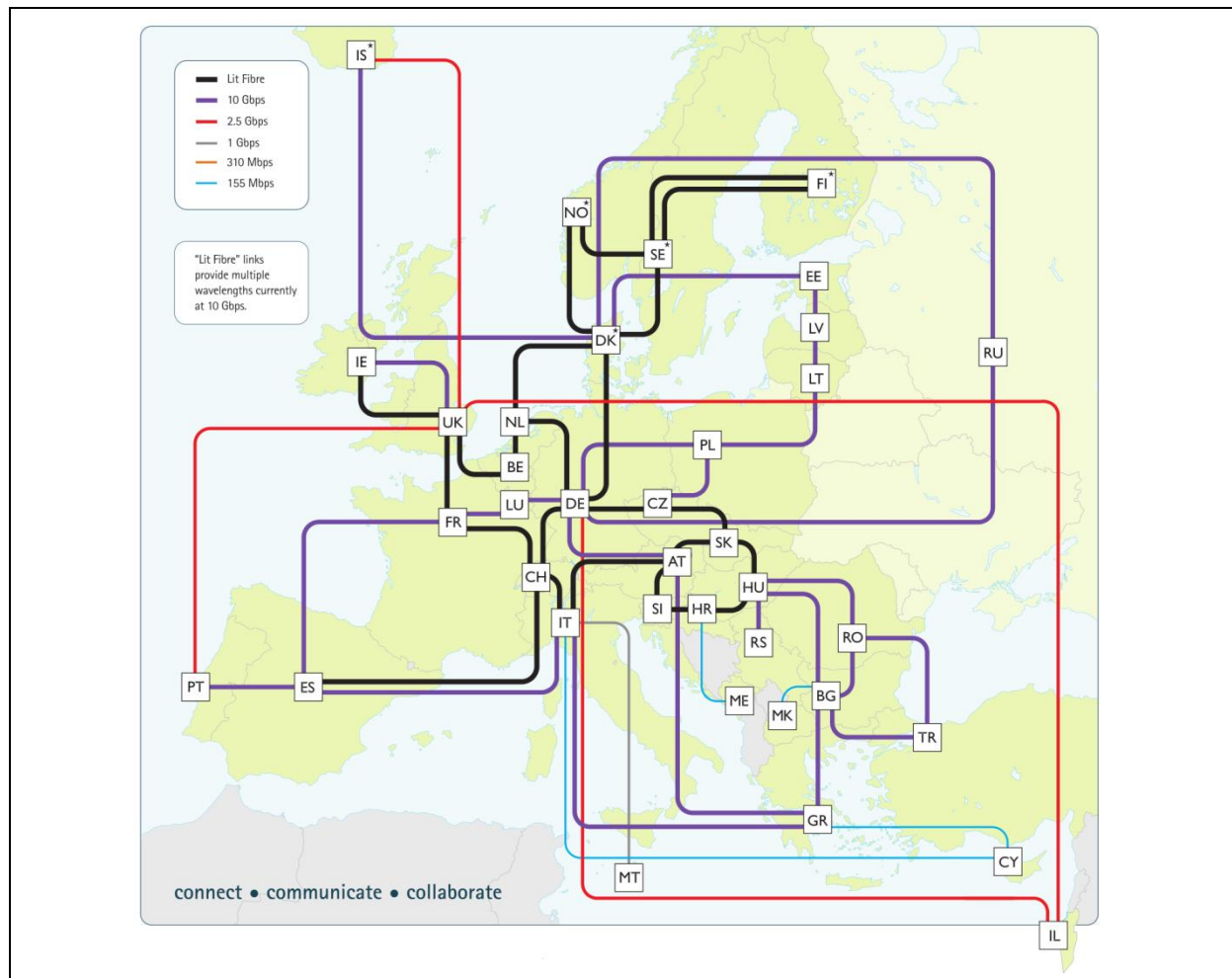


Figure 9 - Topologie GÉANT en Europe - Source GÉANT2.net

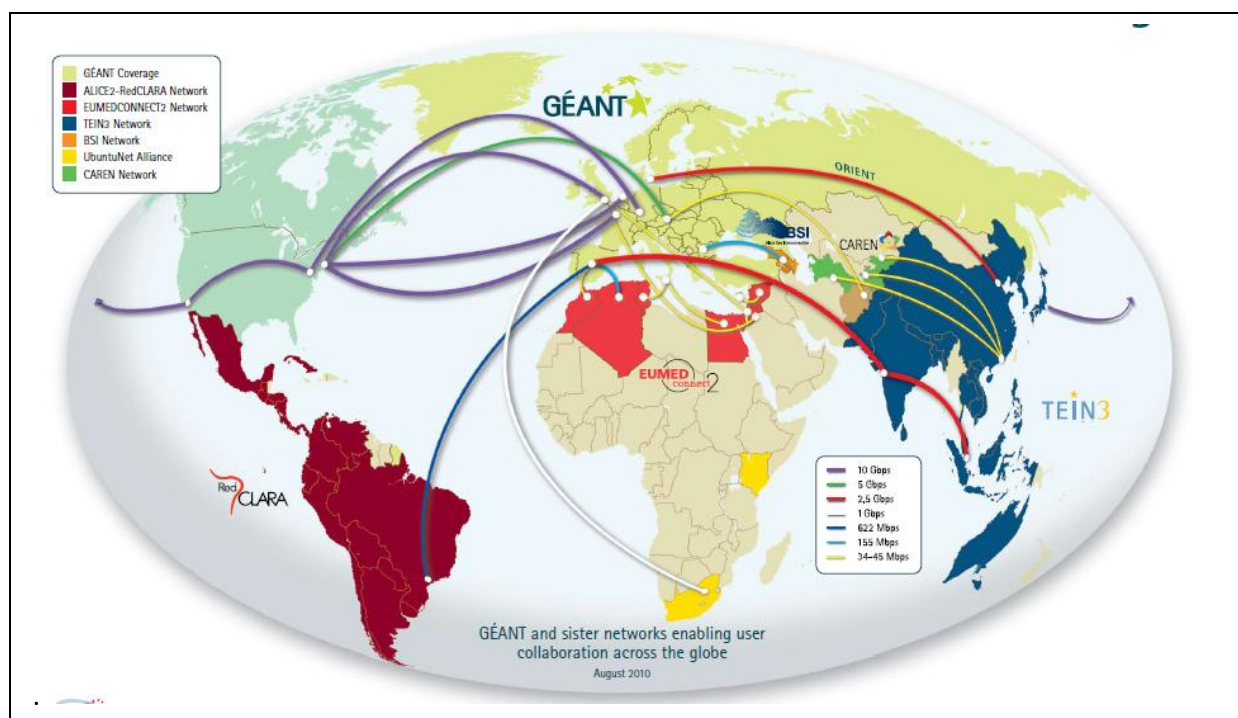


Figure 10 - Réseau GÉANT et débits théoriques mondiaux - Source GÉANT2.net

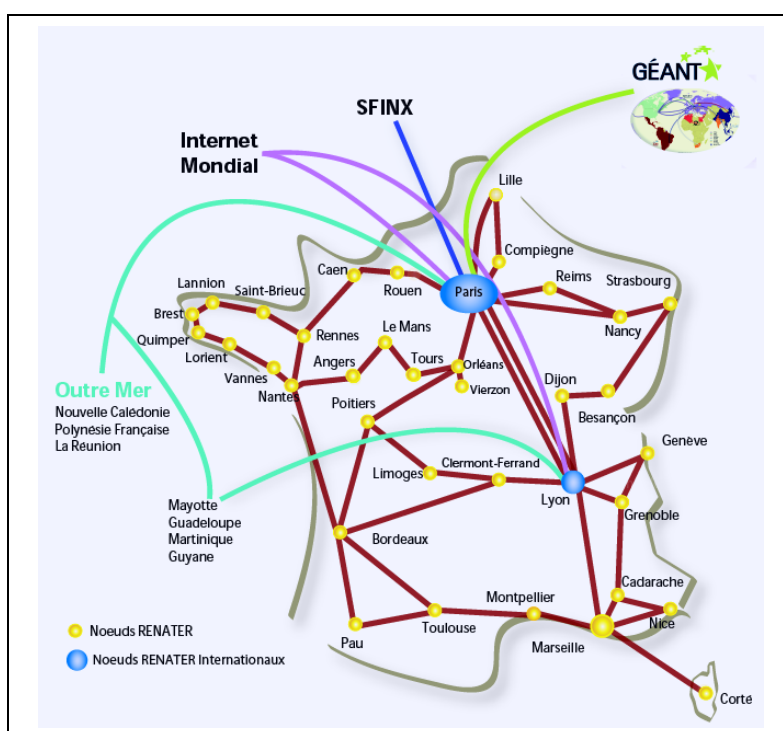


Figure 11 - Réseau Renater - Source Renater.fr

Ces réseaux fonctionnent de façon indépendante d'Internet, où ils sont reliés par des passerelles avec les différents opérateurs. Les avantages d'un tel réseau vis-à-vis d'un opérateur privé sont la neutralité, l'indépendance, la parfaite maîtrise des interconnexions, la gestion poussée de la sécurité et surtout la qualité de service.

Une des principales applications de ces interconnexions est de permettre l'accès distant à tout un ensemble de sites distribués. Par exemple, il est possible d'accéder à un ensemble de données et de les rassembler, virtuellement en un nœud de la grille pour en faire une analyse spécifique. Cette

notion prend tout son sens lorsque l'on se rend compte des difficultés de collecte de l'information, dans ce cas la donnée est non pas collectée mais interrogée directement à distance. En plus de s'affranchir d'un travail de collecte qui, de surcroît ne peut être automatique (ou au mieux semi-automatique), l'accès distant permet d'accéder en temps réel, ou quasi-réel aux données.

Cette fonctionnalité prend une autre dimension lorsque l'on sait que dans le domaine médical, les analyses épidémiologiques sont diffusées au mieux trois ans après la collecte des données. Ces délais sont dus notamment au temps nécessaire à la collecte, la standardisation, le nettoyage, la recherche de doublons, puis à proprement parler l'analyse en elle-même.

Une connexion directe et permanente aux bases de données permettrait de raccourcir drastiquement ces délais à tel point qu'il serait possible d'offrir une surveillance temps réel et de déclencher des alarmes sanitaires avant même qu'un phénomène soit observable sur le terrain (pandémie).

1.3.3.2. Calcul partagé

Avec l'avènement des besoins de calcul et l'essor des super-ordinateurs et des clusters, la plus grande utilité des grilles informatiques est d'effectuer du calcul à large échelle.

Dans ce cas, l'unité de base d'une ressource est le processeur (*CPU*) qui sera alors utilisé par un client via une entité informatique appelé *JOB*. On peut distinguer de grandes familles de ressources, les processeurs des ordinateurs de bureau et les clusters/superordinateurs. Lorsque les *CPU*s sont situés sur un même site, regroupés dans un cluster par exemple, ils sont fusionnés sous l'entité appelée *CE (Computing Element)*.

La principale différence entre un cluster et un superordinateur se trouve au niveau des applications que l'on déploie : un programme superordinateur utilisera l'intégralité du système pendant un temps donné alors qu'un cluster peut exécutera de façon atomique et simultanée un ensemble de programmes. L'architecture des deux systèmes diffère principalement par la vitesse des liens entre les nœuds la composant: un superordinateur disposera souvent de solutions très haut débit et très faible latence, souvent par un système propriétaire là où un cluster utilisera un lien réseau classique de moindre qualité mais autrement plus abordable.

D'un point de vue utilisateur, un superordinateur apparaîtra comme un seul ordinateur exécutant une seule instance d'un système d'exploitation : c'est l'architecture *MIMD*¹. Chaque machine d'un cluster exécutera une instance d'un système d'exploitation.

1.3.3.3. Stockage d'informations

De la même manière qu'une entité *CPU* peut être partagée au sein d'une grille, il est possible de mettre à disposition un élément disposant d'un espace de stockage, avec, comme granularité la plus fine un disque dur ou une capacité de stockage. Ainsi il sera possible à tout utilisateur d'accéder ou de déposer des données situées à travers la grille.

Là où un *CE* rassemble un ensemble de *CPU* d'une même entité géographique, un *SE (Storage Element)* va regrouper un ensemble de ressources de stockage situées au même endroit.

¹ Multiple Instruction stream, Multiple Data stream

Le plus souvent CE et SE sont utilisés de pair afin de fournir en données un job s'exécutant sur un CE (amont) et de lui permettre d'y stocker des résultats (aval). De plus, les volumes de données fournis directement au CE par l'utilisateur en entrée et récupérées en sortie sont de tailles très limitées (quelques Mo), afin de ne pas encombrer ces derniers.

Ces éléments sont donc utilisés à tour de rôle et de façon intense lors de l'exécution de suites de tâches ou workflows comme le montre la [Figure 12].

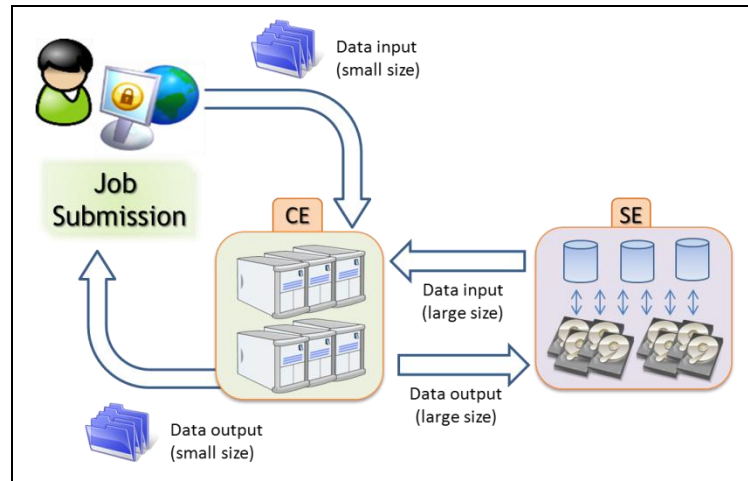


Figure 12 - Interactions CE-SE lors de l'exécution d'un job

1.3.3.4. Autres applications

De la même façon qu'il est possible de partager des ressources de calcul ou de stockage, le concept de grille peut aussi être étendu à d'autres entités, plus hétéroclites, comme des capteurs, de la visualisation, des logiciels ou même des personnes. Libre aux concepteurs d'adapter l'infrastructure à ses besoins.

1.3.4. Les différents types de grilles informatiques

Plusieurs grilles ont été déployées à travers la planète et sont souvent organisées différemment et surtout dédiées à des domaines distincts. On peut considérer tout de même trois types de grilles :

- les grilles dites « de recherche » qui sont dédiées à l'expérimentation académique sur les intergiciels, les systèmes de gestion de données intrinsèques ou des tests de passage à l'échelle. L'utilisateur aura une main mise de très bas niveau, souvent proche du matériel et devra se charger d'un bon nombre d'étapes pour mettre en œuvre ses expérimentations. Ce sont des grilles véritablement dédiées à la recherche informatique et aux informaticiens. Elles permettent de réaliser sans danger les étapes expérimentales indispensables à toute mise en exploitation réelle d'un composant de grille de production ;
- les grilles de production, qui comme leur nom l'indique sont des grilles dédiées à une communauté, souvent issue du monde de la recherche, offrent un outil capable de supporter les exigences en termes de calcul, stockage ou accès à l'information ;
- les grilles privées, qui, sous le terme privé peuvent regrouper deux types d'infrastructures, les grilles administrées et utilisées par une communauté restreinte ou encore le concept des « *desktop grids* », qui visent à utiliser la puissance des ordinateurs personnels à des fins de calcul scientifique ou de partage d'information ;

- le « cloud computing », ou Informatique dans les nuages, peut être considérée comme l'adaptation par le domaine privé des principes des grilles informatiques. Ils sont majoritairement proposés par des grandes entreprises du secteur IT [50] (IBM, Google, Amazon, Salesforce). Le cloud computing permet à tout client de disposer, à la demande, d'un ensemble de ressources informatiques pour effectuer une tâche ponctuelle. La facturation est à la hauteur de la demande, en fonction du temps CPU utilisé, des ressources de stockages, débits et transferts réseaux consommés.
Le « cloud computing » est aussi une réponse à l'«informatique à la demande ». Sa flexibilité permet de réserver les ressources nécessaires au fonctionnement d'une application et ceci autant de temps qu'il sera nécessaire.

1.3.4.1. Les grilles informatiques de recherche

Parmi ces grilles de recherche, on peut citer le projet français de l'INRIA Grid'5000 [51], qui avait pour objectif d'offrir un environnement matériel propice à l'expérimentation de tout type de modèle, intergiciels, systèmes de gestion de données, de sécurité ou encore ordonnanceurs. Ce projet, étendu sur la période 2003-2008 est maintenant repris sous le nom Aladdin¹ pour la période 2008-2012 [52]. Il utilise notamment le réseau Renater pour supporter l'ensemble des besoins de communication.

Dans le même registre, le projet Naregi [53], a permis durant la période 2003-2007 de créer une infrastructure hybride de grille au Japon (recherche/calcul).

1.3.4.2. Les grilles informatiques de production

Le projet EGEE (Enabling Grids for E-science)

Le projet européen EGEE [46] est l'un des plus significatif au monde concernant les grilles de production. Ce projet vise à mutualiser des ressources de calcul et de stockage issues d'universités ou de centres de recherche européens. Le projet, démarré en 2004 suite au projet DataGrid (2002-2004) [54], a permis de créer la plus large grille européenne de calcul et de stockage. L'héritage de ce projet a été confié à EGI [47] qui vise à recentrer au niveau national la gestion de la grille EGEE. L'architecture complète est détaillée en [3.2.1].

Les projets OSG et Teragrid

Open Science Grid [55] est une grille composée essentiellement de clusters situés aux Etats-Unis et en fait la plus grande infrastructure de ce type dans ce pays. Ses missions sont de permettre aux chercheurs et scientifiques d'avoir accès à des ressources de calcul estimées à 37000 CPUs en 2007 [56].

Teragrid [57], est une infrastructure similaire à OSG mais essentiellement constituée de *supercalculateurs*. Les capacités de calcul sont estimées à 2 pétaflops et proposent une quantité de stockage de plus de 50 pétaoctets en 2010 [58].

Plus récemment OSG et Teragrid sont en cours de rapprochement dans une infrastructure commune [59].

¹ A Large-scale Distributed and Deployable Infrastructure

Le projet DEISA

DEISA [60], est une infrastructure Européenne rassemblant tout un ensemble de supercalculateurs issus des centres de recherche. Parmi ces systèmes, on peut citer le JSC¹ qui héberge deux calculateurs dont le plus gros européen du moment en termes de cœurs [61]² avec 294912 unités, l'IDRIS³ du CNRS avec 40000 cœurs, un des plus importants en France, les clusters du CEA au nombre de 3 parmi les 100 plus importants du monde [61] avec un cumul de 161000 cœurs dont le plus puissant suivant les tests, ou encore le BSC⁴, le plus rapide d'Espagne avec 10000 cœurs.

Cet incroyable potentiel de calcul s'est doté d'une infrastructure commune de gestion, c'est le principe moteur de DEISA. Sont concernés principalement l'ordonnancement des tâches, la gestion des données et la sécurité. DEISA propose alors, en réutilisant les intergiciels existants une harmonisation du fonctionnement de ces différents *supercalculateurs*.

L'infrastructure DEISA est particulièrement adaptée à la simulation de processus lourds nécessitant une communication importante entre les différents nœuds de calcul. Une grille de cluster, comme EGEE/EGI, n'est pas du tout adaptée à ce genre d'application car les moyens de communications sont limités au lien Internet d'un site à l'autre.

1.3.4.3. Les grilles privées

On entend par grille privée une utilisation plus restreinte au niveau des utilisateurs ou un mode de financement qui ne dépend pas, ou peu, du domaine public.

Le desktop grid

- Gestion de données :

Parmi les « desktop grid » on peut citer le protocole peer-to-peer Gnutella [62] et ses diverses implémentations logicielles. Il fut le premier à créer un réseau décentralisé d'échange de données mondial. Dans le même créneau on peut citer kazaa [63], edonkey [64] ou plus récemment le protocole bittorrent [65].

- Calcul intensif :

L'application des « desktop grids » pour le calcul intensif est née de deux constats : le besoin grandissant en ressources de calcul et la faible utilisation des CPU des ordinateurs personnels. Ainsi l'initiative Seti@home [66] a vu le jour.

Le projet a deux objectifs : prouver la faisabilité d'un projet de calcul distribué à l'échelle mondiale en utilisant des ordinateurs de bureau principalement et aussi chercher, par l'analyse de signaux extra-terrestres une possible vie en dehors de notre planète. Le premier objectif est considéré comme atteint puisque dès l'année 2002, le projet a accumulé une puissance totale cumulée de 27 Tflops, un record. Concernant le deuxième objectif, il n'a pas été prouvé clairement l'existence de vie extraterrestre mais certaines zones ont montré une activité spectrale qui ne pouvait pas être uniquement due au bruit [67].

¹ Jülich Supercomputing Centre

² Si on occulte le Tera100 du CEA mis en service après la diffusion de cette liste qui 'serait' le plus puissant

³ Institut du Développement des Ressources en Informatique Scientifique

⁴ Barcelona Supercomputing Centre

Seti@home est un des premiers projets issus du programme Boinc [68]. Ainsi, WorldCommunityGrid, FighAids@Home ou Einstein@Home sont des sous-projets issus de Boinc. Sa puissance totale estimée en octobre 2010 est de 3.1TeraFlops [69] avec 300000 utilisateurs actifs, ce qui place Boinc au-delà des meilleurs supercalculateurs du moment [61].

D'autres initiatives ont quant à elles créé leur propre intergiciel indépendamment de Boinc : Condor et XtremWeb [70] ou Folding@home [71]. A noter qu'une tentative d'association entre Boinc, XtremWeb et EGEE/EGI est actuellement à l'étude. Même si le fonctionnement et les objectifs sont différents, une solution a été apportée par EDGeS [72] ce qui prouve que les systèmes ne sont pas totalement incompatibles.

Les grilles informatiques privées

Plus récemment, le concept de grille privée ou plutôt dédiée a fait son apparition. Au sens plus strict du terme que les « desktop grid » les grilles privées sont souvent nées d'une volonté d'une communauté de bénéficier de tout l'avantage des technologies de grille existantes, pour mettre en relation différents protagonistes de leur communauté. Ces avantages liés aux grilles sont de différents domaines : l'évolutivité premièrement, les performances, la sécurité ou encore le volet open-source. Ainsi ces grilles sont très souvent spécialisées et peu étendues, à la différence des grilles de calcul à large échelle qui sont multi-applicatives.

Parmi les grilles strictement privées, c'est-à-dire à usage exclusif on peut citer le projet mammogrid [73] qui vise à créer une infrastructure dédiée à l'analyse de mammographies, le BIRN¹, qui a pour but de fédérer un ensemble de bases de données biomédicales pour la recherche médicale [74] ou encore Astrogrid [75] qui vise à créer une infrastructure dédiée à l'astronomie.

1.3.4.4. Le concept d'organisation virtuelle

Par la suite, le concept d'organisation virtuelle (VO) a facilité la création de « sous-grilles » rassemblant un certain nombre d'utilisateurs et d'institutions qui partagent un ensemble de ressources. Ainsi, Gridchem [76] rassemble une communauté de chimistes, Atlas est une organisation virtuelle d'EGEE dédiée à l'expérience Atlas du Cern [77]. De la même façon, des VO géographiques ont fait leur apparition, souvent à l'échelle d'un pays DGrid [78], d'une région d'un pays : Auvergrid [79], ScotGrid [80] ou d'un ensemble de pays : Ibergrid [81], EuAsiaGrid [82].

Règles de conception des Organisations Virtuelles

La [Figure 13] schématise un ensemble de sites appartenant à une même grille et les différentes organisations virtuelles qui régissent cette infrastructure. Quelques règles sont nécessaires pour assurer la cohérence d'une grille :

- une grille doit posséder au moins 1 nœud ;
- une grille doit posséder au moins 1 VO ;
- une VO peut posséder [1-N] sites ;
- un nœud doit appartenir à [1-N] VO.

¹ Biomedical Informatics Research Network

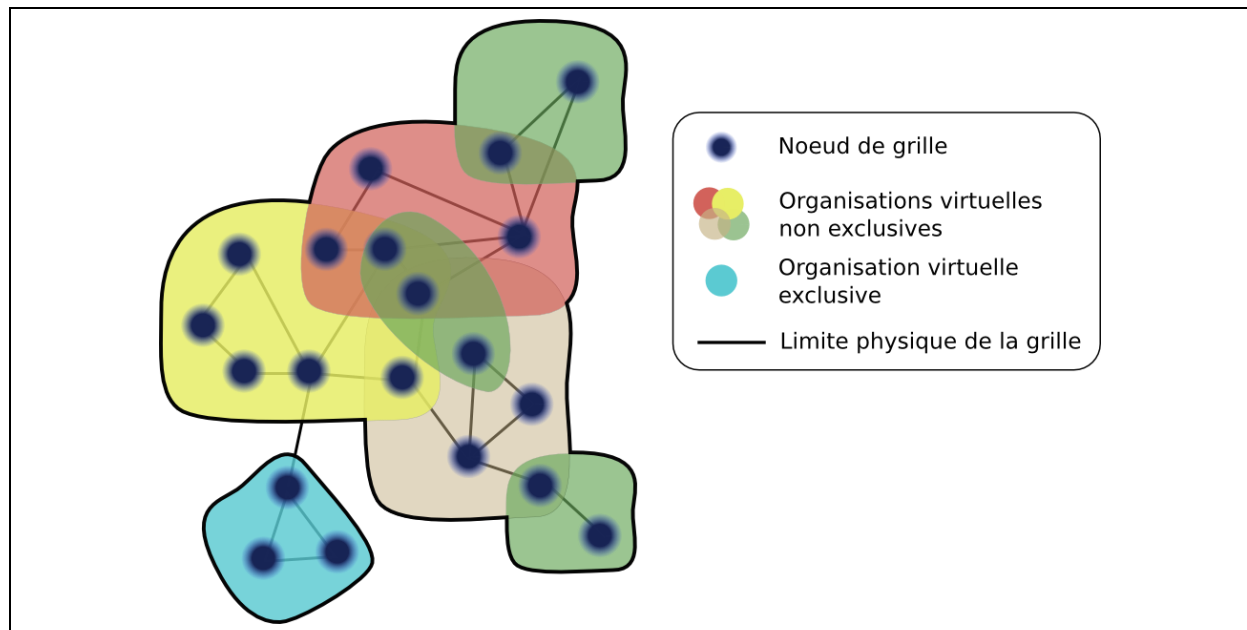


Figure 13 - Représentation schématique des organisations virtuelles

L'organisation virtuelle BIOMED

Au sein de la grille Datagrid, EGEE, puis EGI s'est formée une entité rassemblant les acteurs européens, puis mondiaux du monde médical, biologique et biomédical. La communauté est divisée en trois grands secteurs, l'imagerie médicale, la bioinformatique et la recherche pharmaceutique.

La VO biomed regroupe environ 150 utilisateurs qui se partagent un ensemble de ressources CPU estimées à 17000 cœurs répartis en 100 sites sur 30 pays. A cela s'ajoute une capacité de stockage de 5Pb [Données J.Montagnat 2008].

Cette organisation virtuelle est le berceau de nombreux projets d'utilisation des technologies de grille pour la médecine et a proposé des solutions permettant de tirer parti de ses atouts tout en se conformant aux exigences du monde biomédical.

1.3.5. Les grilles et la protection de la vie privée

Lorsque l'on évoque un environnement de grille, complètement distribué et hétérogène, qui permet tant à tout utilisateur d'accéder, depuis n'importe quel point de la grille, à un ensemble d'informations, il se pose, inévitablement, dans le cas où ces informations sont médicales et de surcroît nominatives un problème crucial de protection de la vie privée.

Les détracteurs potentiels d'un tel système peuvent facilement exploiter cette apparente pervasivité, c'est-à-dire qu'une grille est en proie à une extension naturelle et adaptation dynamique à son environnement qui peut être sous-entendue non-maitrisée, donc manquant de sécurité. Cependant, les concepteurs des grilles ont parallèlement su développer les outils nécessaires et typologiquement adaptés pour assurer cette sécurité, ces aspects seront évoqués ultérieurement dans ce document.

1.3.5.1. Les grilles et le traitement des données médicales

Connaissant les diverses disposition légales à respecter pour faire transiter sur un réseau un ensemble de données sensibles, la question centrale est : comment est-il possible d'adapter une architecture de grille informatique au traitement des données médicales ?

Dans le cadre de l'échange de données médicales utilisant les grilles, J.Herveg, dans un cadre général de l'e-santé [13] puis plus spécifiquement aux grilles informatiques et à l'évaluation des risques [83], a proposé une réponse légale s'appuyant sur la directive européenne 95/46/EC [11].

Malgré le caractère original et novateur des technologies de grilles, il est normal que son manque de diffusion, de notoriété et de popularité soulève des suspicions légitimes de la part des organes décisionnels de la CNIL. Pourtant, à l'échelle auvergnate, certains projets et initiatives comme le PRAI¹ Lifegrid [84], ont tenté de créer des liens entre les acteurs scientifiques du monde (bio)médical et la communauté des grilles informatiques. Bien que ce projet eût l'effet escompté au niveau du rapprochement de ces deux entités scientifiques, les problèmes éthiques, légaux et juridiques n'ont pas été directement évoqués. Pourtant, certains projets issus de Lifegrid ont traité directement de la gestion de l'information médicale [85, 86]. Les raisons de cette absence sont liées à deux domaines : pour le projet Edital [86], celui-ci s'est naturellement affranchi de toute considération légale car le projet était déclaré d'utilité publique (hémovigilance) et hébergé par une structure médicale qui était déjà accréditée pour effectuer ce traitement. Le deuxième projet, HOPE [85] n'a pas abouti à une demande CNIL car il est censé offrir un outil dédié à la télémédecine en utilisant les technologies de grille, il ne devait pas se charger de son implémentation à l'intérieur de structures hospitalières, ce qui a quelque peu nui à son essor. Pourtant, ce projet implémentait l'ensemble des recommandations nécessaires à la certification CNIL, notamment une encryption point à point de toutes les données transitant sur le réseau, un stockage sûr des informations dans les bases de données locales et avec réplication sous forme chiffrée sur la grille, ou encore une authentification forte en utilisant la couche PKI² des grilles informatiques. De la même manière, le projet ANR NeuroLOG [87], respectait aussi toutes ces recommandations [88] sans pour autant avoir demandé une accréditation CNIL.

Certains projets orientés gestion de données médicales sur grille ont pourtant réussi à obtenir leur accréditation CNIL comme Health-e-child [89] ou Mammogrid [73] mais restent des exceptions.

1.3.6. Transposition pour le dépistage des cancers en Auvergne

1.3.6.1. Expériences

L'utilisation des grilles privées a montré son efficacité pour aborder le problème de l'accès aux données médicales distribuées. Les projets Mammogrid [73] et eDiamond [90] ont été les précurseurs d'un tel système de gestion de données. Par la suite la communauté HealthGrid s'est créée, visant à rassembler les acteurs du « *grid computing* » et les acteurs du monde médical. Le papier blanc de

¹ Programme Régional d'Actions Innovatrices

² Public Key Infrastructure

l'association Healthgrid [91, 92], a posé de nombreuses bases et servi de référence à de futurs travaux dans ce sens :

“Healthgrids are Grid infrastructures comprising applications, services or middleware components that deal with the specific problems arising in the processing of biomedical data. Resources in Healthgrids are databases, computing power, medical expertise and even medical devices. Healthgrids are thus closely related to eHealth”

C'est sur ces foundations que d'autres initiatives telles qu'ACGT [93] ou Health-e-child [89] ont amélioré le concept pour rendre plus interopérable et accessible au personnel médical le monde de la grille.

Ces différentes expériences ont permis de faire une première preuve de concept en amenant pour certaines d'entre elles « la grille à l'hôpital ». C'est dans cette optique que peut s'insérer la problématique du dépistage des cancers en Auvergne.

1.3.6.2. Méthodologie

Le dépistage des cancers en Auvergne se heurte à deux problèmes majeurs :

- l'accès aux données médicales par les structures de dépistage organisé d'une part ;
- la réticence des professionnels de l'anatomie pathologie à partager leurs données médicales d'autre part et cela pour des raisons liées à la propriété et la sécurité des données.

En effet, les anatomo-pathologistes ont un sentiment de « vol » des données lorsqu'elles sont demandées par les hautes instances nationales de l'épidémiologie. Ainsi, le syndicat des pathologistes a recommandé à toute la profession de refuser de façon systématique l'envoi de données en dehors de leur structure tant que cette situation perdurerait. De plus, la collecte des données se faisant essentiellement par un système déclaratif, le pathologiste prend de son temps de travail pour effectuer une déclaration sans pour autant percevoir une quelconque reconnaissance, qu'elle soit financière ou simplement une mention dans les résultats et exploitations épidémiologiques.

Système déclaratif / système connecté

Le système historique de collecte de données se base sur une déclaration du médecin, en réponse à la demande d'un client comme montré en [Figure 14], qui ici est par exemple l'InVS¹.

L'inconvénient majeur d'un tel système est qu'il implique le médecin dans la chaîne de collecte des données, l'obligeant à recueillir l'information dans ses bases de données, la transformer, la nettoyer et la présenter de façon claire pour l'envoyer, souvent de façon archaïque à un client qui en fait la demande. Connaissant l'emploi du temps des praticiens et le manque de reconnaissance par les organismes effectuant la demande, il est normal que celui-ci émette une certaine réticence à effectuer un travail qui n'est pas a fortiori le sien.

¹ Institut National de Veille Sanitaire

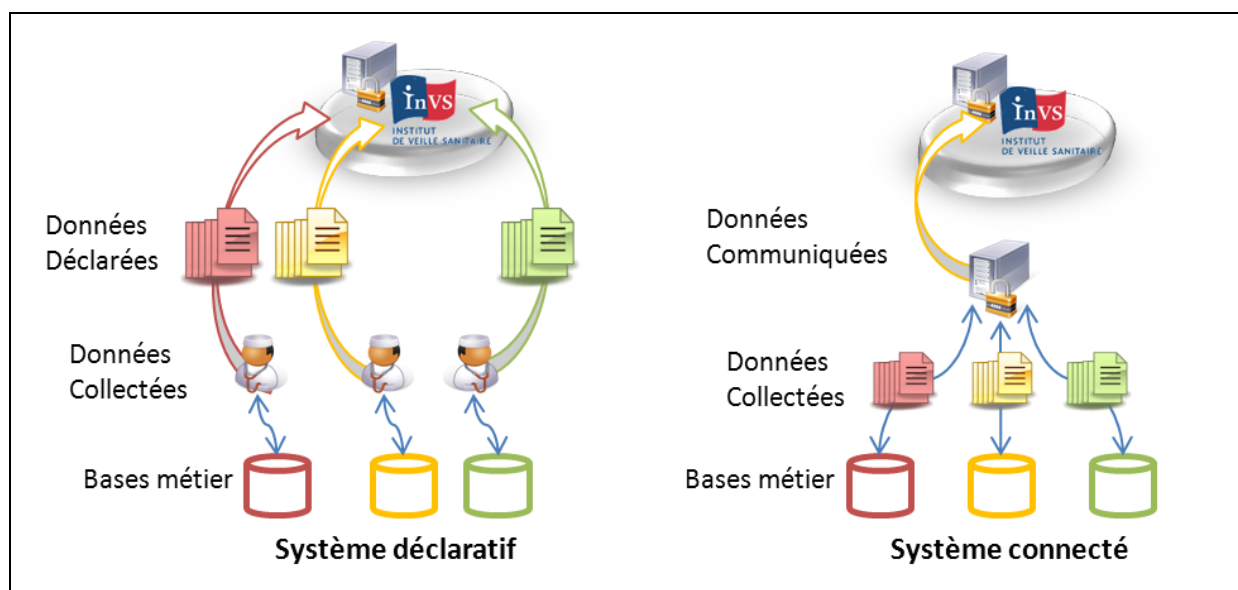


Figure 14 - Système déclaratif / système connecté

D'autre part, avec les moyens informatiques et de communication actuels, une autre méthode de recueil des données fait son apparition, le système connecté, qui affranchit le médecin de cette étape en interrogeant directement ses propres bases. L'autre avantage est bien entendu la rapidité d'accès et de disponibilité des données, puisqu'elle ne dépend pas d'une intervention humaine. De plus, les méthodes de recueil peuvent être harmonisées dans toutes les structures médicales ce qui, économiquement parlant, est très appréciable.

Malgré l'apparente « perte de contrôle » d'un système connecté, il est facilement démontrable qu'il s'agit de l'exact opposé, puisque le client des données « puise » directement dans les bases sans en extraire massivement le contenu. Pour un système déclaratif : « ce qui est envoyé est perdu » aucune modification ni rectification ne peut se faire post-envoi sans un protocole manuel.

Système connecté en utilisant les grilles informatiques

Ainsi, l'implantation d'une grille reliant les différents protagonistes du dépistage organisé est envisageable dans le cadre de la création d'un réseau d'échange de données. Il doit pour cela se conformer aux différentes exigences des différents acteurs du réseau, en apportant les garanties nécessaires à leur consentement.

Il doit aussi se conformer aux différentes dispositions légales présentées dans ce chapitre, afin de garantir toute la sécurité et d'assurer la confidentialité des données médicales.

L'étape indispensable de tout développement de ce type est de concevoir un cahier des charges qui comportera une spécification détaillée des acteurs, des besoins et des contraintes d'implémentation.

CONCLUSION

Ce chapitre présente dans un premier temps et de façon globale le fonctionnement de l'e-santé en France et soulève les problèmes techniques, éthiques et légaux que rencontre toute la chaîne de traitement des données dans le cadre du dépistage organisé des cancers.

L'obligation du respect des lois informatique et libertés au niveau national et de la directive européenne 95/46/EC [11], qui visent à protéger les patients du traitement de leurs données de santé, imposent de nombreuses contraintes à la création d'un réseau d'échange de données cancer. Principalement, on peut retenir deux familles de mesures à prendre en compte :

- l'information : il faut s'assurer que tout patient puisse connaître l'existence d'une entité gérant ses données par un moyen direct et surtout lui donner le total accès pour toute opération de modification ou d'opposition ;
- la protection : les données doivent respecter un ensemble de mesures nécessaires à garantir la sécurité des informations. Toute activité sur l'infrastructure doit en plus être tracée, afin de pouvoir remonter à l'origine d'une éventuelle fuite ou corruption de données.

Une étape de validation du respect de ses dispositions est nécessaire par le biais d'une demande d'autorisation auprès de la CNIL.

Dans un deuxième temps sont présentées les grilles informatiques et les technologies associées et comment elles peuvent proposer une piste de réponse tangible pour permettre d'améliorer l'efficacité du dépistage organisé en Auvergne. Par ailleurs, les avantages d'un tel système sont montrés, l'accent est porté sur les améliorations en ce qui concerne l'accès aux données et la façon de les mettre à disposition. Les contraintes concernant le respect de la vie privée et des différents textes de lois l'encadrant ne sont pas oubliées et la faisabilité d'un système propulsé par une grille informatique est montrée, notamment en s'appuyant sur des projets pilotes existants.

De plus, l'équipement dont bénéficie l'Auvergne, grâce en particulier au projet Auvergrid [79] et au réseau Auverdata qui lui est associé, prédispose la région, en termes d'infrastructure comme de compétences à l'utilisation de ces technologies.

Chapitre 2. Cahier des charges du projet RSCA

INTRODUCTION

La deuxième phase de ce document concerne le projet RSCA : Réseau Sentinelle Cancer Auvergne. Il constitue la base applicative du travail de thèse. Il a pour objectif principal de permettre la mise en relation de données cancer sur la région et cela à des fins d'échange de données et d'études épidémiologiques.

Ce chapitre présente ainsi le cahier des charges du projet RSCA. Son écriture fut l'occasion de spécifier les objectifs, besoins et contraintes du projet, de rencontrer les différents acteurs et de décrire le fonctionnement du système.

Son écriture a été effectuée lors de la première année de thèse en collaboration avec l'ensemble des acteurs et il a été volontairement laissé en l'état de sa version finale, c'est-à-dire celle validée par l'ensemble du consortium en novembre 2008, lorsqu'a été créée l'association *RSCA*. Cette association a pour objectif de fournir une structure juridique d'accueil, de gestion et d'administration du projet. Elle a aussi pour but d'assurer l'interface et la cohésion entre les partenaires médicaux participant au projet.

Le cahier des charges a aussi permis d'offrir un document support pour toute décision administrative ou demande de subvention afin d'assurer financièrement le lancement du projet.

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	1 sur 38

Réseau Sentinelle Cancer Auvergne

CAHIER DES CHARGES FONCTIONNEL

Auteurs :

Pierre Bouchet
Paul De Vlieger
Alain Gaillot
Marie-Ange Grondin
Lydia Maigne
Chantal Mestre
Pâquerette Lonchambon
Josette Puvinel
TianXiao Wei

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	2 sur 38

Sommaire

SOMMAIRE	2
TABLE DES ILLUSTRATIONS.....	4
1 INTRODUCTION.....	5
1.1 PRINCIPES FONDATEURS.....	5
1.2 DESCRIPTION DU PROJET	6
1.3 INTERET ET OBJECTIFS GENERAUX DU PROJET.....	7
2 ACTEURS.....	8
2.1 LES ACTEURS/PARTENAIRES CONCERNES	8
2.2 LES ACTEURS DU PROJET "RESEAU SENTINELLE"	9
2.2.1 Associations	9
2.2.2 Cabinets et services d'Anatomie et de cytologie -pathologique	10
2.2.3 Les structures de santé publique.....	11
3 ANALYSE	13
3.1 FONCTIONNEMENT ACTUEL	13
3.1.1 Pour les associations :	13
3.1.2 Pour les cabinets d'anatomo-pathologie	13
3.1.3 Pour la santé publique	13
3.1.4 Contraintes.....	13
3.2 ANALYSE DE L'EXISTANT	14
3.2.1 Logiciels Métier	14
3.3 OBJECTIF	15
3.3.1 Dépistage par les associations	15
3.3.2 Indicateurs évaluatifs.....	17
3.3.3 Scientifique.....	17
3.4 PERSPECTIVES (OBJECTIF)	17
3.4.1 A court terme	18
3.4.2 A plus long terme	18
3.5 MOYENS DU PROJET.....	19
3.6 BENEFICES ATTENDUS	19
3.6.1 Pour les associations :	19
3.6.2 Pour les structures de santé publique	19
3.6.3 Pour les cabinets d'anatomo-pathologie	20
3.7 LISTE DES OBJECTIFS :	20
3.7.1 Fonctionnels	20
3.7.2 Techniques	21
3.7.3 Administration et maintenance	21

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	3 sur 38

	3.7.4	Juridiques	21
3.8		PERSPECTIVES DE SOLUTION	22
4		DESCRIPTION FONCTIONNELLE.....	23
4.1		CARACTERISTIQUES ET FONCTIONNEMENT DU SYSTEME.....	23
	4.1.1	Description des parties client	23
4.2		SCENARIOS D'UTILISATION DU SYSTEME PAR ACTEUR.....	24
	4.2.1	Scénario d'utilisation par les anatomo-pathologistes	24
	4.2.2	Scénario d'utilisation par les administrateurs	25
	4.2.3	Scénarios d'utilisation par les associations de dépistage organisé	26
	4.2.4	Scénarios d'utilisation par les structures de santé publique	27
	4.2.5	Imagerie médicale.....	27
5		DONNEES.....	28
5.1		TYPES DE DONNEES	28
	5.1.1	Données patients	28
	5.1.2	ADICAP	28
	5.1.3	Images médicales (Mammographies...).....	28
5.2		DONNEES PAR ACTEUR	28
	5.2.1	Associations de dépistage organisé.....	28
	5.2.2	Pathologistes (descriptifs des fiches de comptes-rendus standardisés)	29
	5.2.3	Statistiques/Epidémiologie	29
6		SECURITE	31
6.1		IDENTIFICATION UTILISATEUR.....	31
	6.1.1	Carte CPS.....	31
	6.1.2	Gestion des habilitations.....	31
6.2		IDENTIFICATION PATIENT	32
7		ECHEANCIERS (MISE EN ŒUVRE).....	33
8		VALIDATION.....	34
9		ANNEXES.....	35
9.1		BIBLE DES DONNEES ANATOMO-PATHOLOGIQUES STANDARDISEE	35
9.2		GLOSSAIRE (VOCABULAIRE)	36
10		CREDITS	37
10.1		ACTEURS TECHNIQUES	37
10.2		AUTRES ACTEURS	37
10.3		ANATOMO-PATHOLOGISTES	37
10.4		ASSOCIATIONS.....	38
10.5		EPIDEMIOLOGIE	38

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	4 sur 38

Table des illustrations

Figure 1 -	Objectif Associations	15
Figure 2 -	Objectif épidémiologie	16
Figure 3 -	Fonctionnement du système	23
Figure 4 -	Cas d'utilisation pour les pathologistes	24
Figure 5 -	Scénario d'utilisation Administrateurs	25
Figure 6 -	Scénario d'utilisation pour les associations	26
Figure 7 -	Scénario d'utilisation pour les acteurs de santé publique.....	27
Figure 8 -	Données nécessaires à l'épidémiologie	30
Figure 9 -	Exemple d'interface de gestion des droits	32

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 1 : INTRODUCTION	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	5 sur 38

1 Introduction

Cette partie présente le projet qui sera communément appelé « réseau sentinelle » dans le reste du document et portera sur le dépistage organisé du cancer en Auvergne et ses applications en santé publique.

1.1 *Principes fondateurs*

1^{er} principe

Ce projet s'inscrit dans la problématique globale de la lutte contre le cancer en Auvergne. Depuis 1994 pour l'Allier et 1995 pour le Puy-de-Dôme, les campagnes de dépistage organisées du cancer du sein ont été mises en place. Depuis 2003, l'ensemble de la population féminine âgée de 50 à 74 ans de la région Auvergne bénéficie d'une offre de dépistage gratuit par mammographie assuré par les deux associations chargées de ce dépistage organisé : l'ABIDEC¹ et l'ARDOC². Ces deux associations ont ensuite développé à partir de 2004 le dépistage organisé du cancer colorectal, pour les hommes et les femmes de 50 à 75 ans, sur l'ensemble du territoire régional.

Afin de mener à bien, et de façon la plus efficace possible les missions dont ces associations ont été chargées, elles doivent réaliser un recueil d'information auprès des différents acteurs impliqués : anatomo-pathologistes, médecins traitants, radiologues voire le patient. Ce véritable travail de fourmi pourrait grandement être facilité en mettant en relation les données "patient" détenues par ces acteurs.

A l'heure actuelle, ce recueil de données se fait manuellement, par du personnel chargé de la ressaisie informatique des données anatomo-pathologiques. Ces données existent pourtant déjà sur support électronique depuis plusieurs années.

Devant les difficultés rencontrées et pour rendre l'ensemble du système plus réactif, est né le projet de créer un réseau de communication pour mettre à disposition des clients potentiels les différentes sources d'information. Ce réseau sentinelle pourrait être interrogé par des acteurs du domaine médical ou scientifique qui seraient intéressées par ces données, notamment dans la santé publique.

¹ ABIDEC = association bourbonnaise interdépartementale du dépistage des cancers

² ARDOC = association régionale de dépistage organisé des cancers

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 1 : INTRODUCTION	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	6 sur 38

2eme principe

Ce projet doit s'inscrire dans le cadre général de la santé publique : la santé publique s'intéresse à la santé des populations en développant une approche globale des questions de santé conformément à la définition adoptée par l'OMS (la santé comme un état de bien-être complet, physique, mental et social) et aux stratégies de promotion de la santé.

C'est un ensemble des connaissances et techniques relatives à l'état de santé d'une population, aux déterminants collectifs de santé et aux techniques propres à améliorer la santé du plus grand nombre. La population ciblée peut être l'ensemble des personnes d'une zone géographique déterminée (pays, région, département, canton, ville, quartier...), ou d'un groupe présentant une caractéristique commune (âge, handicap, situation sociale...). Les outils d'observation et d'analyse sont la démographie, l'épidémiologie, la sociologie, la psychosociologie, l'anthropologie.

Mais c'est aussi un ensemble d'actions et prescriptions prises par l'administration, de programmes d'intervention au plus près des citoyens.

3eme principe

Dans une première étape, pour des raisons pratiques, ce projet doit se limiter à la lutte contre le cancer en liaison avec les organisations de dépistages. En conséquence, le présent cahier des charges définit les conditions de réalisation et de fonctionnement du réseau sentinelle.

1.2 Description du projet

Ce document s'organise de la façon suivante :

La partie [2 : Acteurs p.8] présente les différents acteurs du projet et fait la description des principaux : les associations de dépistage organisé des cancers, les cabinets d'anatomo-pathologie et les structures de santé publique. Ces présentations sont accompagnées d'une estimation chiffrée de leur activité.

Ensuite, vient une phase [3 : Analyse p.13] de description et d'analyse du fonctionnement actuel, on donnera ici le point de vue de la santé publique et du dépistage organisé. En découlent les objectifs et perspectives d'objectifs du projet, ses moyens et bénéfices attendus.

Après cette description des objectifs une description fonctionnelle est faite [4 : Description fonctionnelle p.23], avec les différents scénarios d'utilisation pour les acteurs du réseau sentinelle. La partie suivante [5 : Données p.28] présente la description des données qui seront échangées. Tout ceci sera régi par des contraintes en matière de sécurité et de protection des données [6 : Sécurité p.31].

Ensuite un échéancier et une partie introductive sur la future validation des travaux sont présentés.

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 1 : INTRODUCTION	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	7 sur 38

1.3 Intérêt et objectifs généraux du projet

Au démarrage du projet, les objectifs principaux et directs sont :

Pour les associations de dépistage :

- permettre de réunir ou rendre disponible les données nécessaires à l'évaluation de l'efficacité des dépistages (taux de détection des cancers) ;
- remplacer la collecte manuelle de données anatomo-pathologiques par un système informatisé ;
- a une plus grande échelle, pouvoir obtenir de manière fiable des données épidémiologiques sur les cancers traités en région.

Pour les structures de santé publique :

- obtenir de manière plus exhaustive des informations sur l'ensemble des cancers dans la région Auvergne ;
- rendre disponible plus rapidement et avec une plus grande précision ces informations afin d'améliorer la pertinence des analyses épidémiologiques sur la région et la réactivité des autorités.

Si cette première phase est menée à bien pour les associations de dépistage et la santé publique, il sera techniquement possible dans l'avenir :

- de positionner ce réseau sentinelle comme centre d'une plateforme Auvergne de santé publique ;
- de rassembler les différents acteurs du domaine de la cancérologie ou encore médical ;
- de posséder une structure juridique d'accueil, de maintenance et d'utilisation du réseau pour sa pérennisation.

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 2 : ACTEURS	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	8 sur 38

2 Acteurs

Le réseau sentinelle fonctionnera avec les acteurs majeurs de la santé publique sous l'égide d'une structure de gestion. (voir partie [3.5 :Moyens du projet p.19])

2.1 *Les acteurs/partenaires concernés*

Deux groupes d'acteurs sont à distinguer :

Les acteurs dits « primaires » du projet. Ils appartiennent à 3 grands domaines :

- les utilisateurs du réseau sentinelle, à savoir les associations chargées du dépistage organisé des cancers du sein et du colon ;
- les fournisseurs de données, c'est-à-dire les différentes structures anatomo-pathologiques de la région ;
- les développeurs du projet, chargés de l'implémentation, du déploiement et des tests.

Les acteurs dits « secondaires » seront inclus dans le projet une fois la structure informatique et juridique mise en place. Des utilisateurs ou fournisseurs pourraient rejoindre le réseau. Certains utilisateurs pourraient à leur tour devenir fournisseurs de leurs propres données. Tous ces échanges seront réglementés par une charte.

Acteurs primaires:

- Ces associations sont au nombre de deux dans la région Auvergne, l'ABIDEC pour le département de l'Allier et l'ARDOC pour les 3 autres départements de l'Auvergne : le Cantal, la Haute-Loire et le Puy-de-Dôme.
- Les fournisseurs de données : principalement les anatomo-pathologistes de la région, situés dans les centres de lutte contre le cancer, les hôpitaux et les cabinets. Ces fournisseurs seront présentés en détail par la suite.
- Les développeurs :
 - Le Laboratoire de Physique Corpusculaire (LPC): Equipe (Plate-forme de Calcul pour les Sciences du Vivant)
 - L'Equipe de Recherche en Imagerie Médicale (ERIM)
 - Le Service de Santé Publique du CHU de Clermont-Ferrand, qui aura aussi un rôle de client pour des requêtes statistiques.

Acteurs secondaires : (second temps)

Des acteurs supplémentaires pourraient intégrer le réseau, que ce soit pour l'enrichissement du réseau sentinelle ou en tant que clients des données :

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 2 : ACTEURS	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	9 sur 38

- en fournisseur apparaîtront les cabinets de radiologie, qui pourraient contribuer à un échange informatisé d'images radiographiques (mammographies)
- les cabinets de gastroentérologie (publics et privés) (nécessaire au cancer du colon)
- en utilisateurs pour l'épidémiologie pourraient figurer notamment des services de l'Etat (CIRE, DRASS, future ARS (Agence Régionale de Santé)).

2.2 Les acteurs du projet "réseau sentinelle"

2.2.1 Associations

L'organisation des dépistages a été confiée par l'Etat à des associations départementales. Ces associations fonctionnent grâce à un financement public, partagé entre l'Etat et l'assurance maladie.

Deux dépistages sont, à l'heure actuelle, mis en œuvre en région Auvergne : celui du sein et celui du colon. A terme, le dépistage organisé pourra s'étendre à d'autres types de cancers, notamment les cancers dits évitables comme le cancer du col de l'utérus (en expérimentations dans certaines régions), celui de la prostate, ou des mélanomes.

Les deux associations concernées ci-après seront les premiers clients des données mises à disposition par le réseau sentinelle.

Leur mission est d'inviter les personnes éligibles à effectuer leur dépistage dans le cadre du Dépistage Organisé et d'assurer leur suivi en cas de test positif. Pour cela, ces associations doivent recueillir des données histologiques sur ces personnes.

L'ARDOC

L'ARDOC, « Association Régionale de Dépistage Organisé des Cancers » est une association chargée d'assurer le dépistage des cancers du sein et du colon dans les départements du Puy-de-Dôme, du Cantal et de la Haute-Loire.

Chiffres de référence 2007

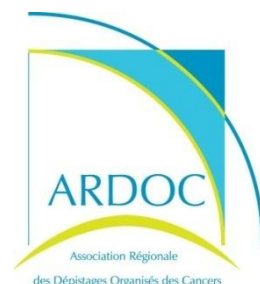
Pour le dépistage du cancer du sein :

- 180 dossiers/jour à saisir -> 2 lectures de mammographies par professionnels à faire
- 40000 dossiers/an -> 600 relectures de comptes rendus anatomopathologiques

Pour le dépistage du cancer du colon :

- 200 dossiers/jour à saisir
- 45000 dossiers/an -> 1400 relectures de comptes rendus anatomopathologiques

A titre d'exemple : pour l'ARDOC (pour 3 départements):



12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 2 : ACTEURS	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	10 sur 38

Si l'on considère uniquement le cancer du sein c'est environ 1500 femmes avec mammographies classée ACR¹ 3-4-5 (positif) qui doivent être suivies soit 7 h par semaine pour les travaux de relance, de collecte, de saisie et traitement dossier par le personnel administratif et médecin coordonnateur. Pour le cancer colorectal, l'activité est double par rapport au sein.

L'ABIDEC

L'ABIDEC, « Association Bourbonnaise Interdépartementale du Dépistage des Cancers » opère de la même façon que l'ARDOC mais dans le département de l'Allier.



Chiffres de référence 2007

Pour le dépistage du cancer du sein :

- 15700 dossier/mammographies
- 600 positifs -> 370 relecture de comptes rendus anatomopathologiques

Pour le dépistage du cancer du colon :

- 23000 tests lus
- 667 positifs -> 350 comptes rendus anatomo-pathologiques relus

2.2.2 Cabinets et services d'Anatomie et de cytologie - pathologique

Métier

Ils sont chargés d'examiner les pièces opératoires et les biopsies prélevées par différentes spécialités médicales et chirurgicales : chirurgiens généraux ou spécialisés, gynécologues, gastro-entérologues, ORL, Stomatologues, dermatologues, internistes, endocrinologues, médecins nucléaires, hématologues, radiologues, voire généralistes.

Ils effectuent tous les diagnostics de cancers sans exception ; fournissent les facteurs pronostics et prédictifs (de réponses aux traitements). Leurs données, leurs rapports d'expertise sur les pièces examinées permettent de couvrir l'ensemble de la cancérologie en France. En effet, il n'y a pas de cancer traité sans compte rendu anatomo-pathologique d'où l'intérêt porté par les associations de dépistage et la santé publique (voir ci-après).

Ils effectuent d'autres diagnostics encore, en particulier pour les lésions inflammatoires (peau, système digestif, pathologie ganglionnaire...)

¹ American College of Radiology : Classifications des anomalies mammographiques

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 2 : ACTEURS	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	11 sur 38

Liste des structures incluses dans le projet

Cette liste contient l'ensemble de l'activité d'anatomo-pathologie de la région.

- Centre de lutte contre le cancer Jean Perrin – Clermont-Ferrand
- Service du Pr. Kémény – CHU Gabriel Montpied
- Service du Pr. Déchelotte – CHU Hôtel Dieu
- Cabinet Sipath (Clermont-Ferrand)
- Dr. Cloup-trochon – Montluçon

2.2.3 Les structures de santé publique

Définition

La santé publique est l'art d'utiliser la science dans un contexte politique, social, économique, afin de réduire les inégalités en matière de santé, tout en assurant la meilleure santé possible au plus grand nombre (Organisation Mondiale de la Santé 1998). Les champs d'intervention sont extrêmement variés : vaccinations, la lutte contre les fléaux sociaux, l'éducation pour la santé, le contrôle des établissements sanitaires, médico-sociaux, des pharmacies, des laboratoires ou des organismes d'assurance maladie, la santé environnementale : air, eau, habitat..., les aliments (OGM...), l'industrialisation (pollution...), voire la protection civile (terrorisme et plans d'urgence).

Le service de santé publique du CHU de Clermont-Ferrand

Le service de Santé Publique du CHU a une double vocation : hospitalière et universitaire avec la recherche scientifique d'une part, acteur de la santé publique pour la région d'autre part. De ce fait, il participe au développement de réseaux sentinelle des cancers selon l'angle de la recherche scientifique (notamment quant à la validité des pratiques de veille sanitaire et aux analyses sensibilité/spécificité des systèmes de surveillance et d'alerte) et de l'analyse institutionnelle, ainsi qu'à son exploitation en termes d'information des décideurs, en partenariat avec les instances régionales concernées. Cette information concerne, entre autres, les indicateurs régionaux de veille sanitaire et de diagnostic de santé publique nécessaires à la prise de décision et à la recherche en épidémiologie, économie de santé et prévention.

Le service de santé publique s'intéresse au projet sous deux aspects :

- la veille sanitaire qui vise à répondre aux questions d'épidémiologie générale des cancers et à celle de l'étude des cas-groupés de maladie
- l'évaluation des politiques de santé publique en termes de prévention primaire (réduction de l'incidence par réduction des risques) et secondaire (réduction de la prévalence par le dépistage)

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 2 : ACTEURS	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	12 sur 38

Il assurera en particulier l'évaluation du réseau sentinelle cancer auvergne: analyse de la sensibilité/spécificité du système de surveillance et d'alerte, évaluation de ses résultats en termes d'informations et de partenariats.

Personnes : Pr. Laurent Gerbaud/ Dr. Lemlih Ouchchane / Dr. Marie-Ange Grondin/ C Auclair

Le groupement régional de santé publique (GRSP)

Créé par la loi d'août 2004, les groupements régionaux de santé publique ont pour mission la mise en œuvre des plans régionaux de santé publique (PRSP). Dotés d'un conseil d'administration, présidé par le Préfet de Région, ils réunissent au sein d'un groupement d'intérêt public (GIP) l'ensemble des acteurs concernés par la santé publique, y compris les collectivités territoriales volontaires.

Le GRSP Auvergne comporte un programme de lutte contre le cancer dont un des objectifs est de développer les dépistages des cancers, d'en assurer le suivi et l'évaluation au plus près des réalités de la région.

Le Préfet de région a posé la candidature de l'Auvergne pour participer à l'expérimentation du système multi-source de surveillance des cancers porté par l'InVS. Les trois sources retenues sont les données ALD (affections de longue durée), PMSI (programme médicalisé des systèmes d'information) et précisément les données anatomopathologiques. Le projet de réseau sentinelle, dans sa partie santé publique s'inscrit dans la démarche du futur système de l'InVS.

Compte tenu des réformes d'organisation de l'administration de la santé en région que devrait entraîner la loi « hôpital, patients, santé et territoire », les compétences actuelles du GRSP devront être transférées à l'agence régionale de santé (ARS) en 2010.

CIRE : Cellule Inter Régionale d'Epidémiologie

La CIRE est l'acteur de la région de l'InVS, lequel est chargé notamment de la surveillance nationale des cancers, du suivi épidémiologique des dépistages organisés ; l'InVS participe par ailleurs au système d'alertes sanitaire et apporte sa contribution à la gestion des situations exceptionnelles.

OBRESA : Observatoire régional de la santé d'Auvergne

Il est en charge de l'observation sanitaire régionale. Actuellement l'OBRESA gère le réseau de surveillance des mélanomes, sachant que l'un des objectifs à moyen terme est d'étendre ce projet à la surveillance des cancers. Sur le plan national, l'OBRESA est membre de la FNORS (fédération nationale des observatoires régionaux de la santé).

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 3 : ANALYSE	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	13 sur 38

3 Analyse

3.1 *Fonctionnement actuel*

3.1.1 Pour les associations :

Les associations qui organisent le dépistage des cancers du sein et du colon sont chargées dans le cas de positivité chez les sujets dépistés, du suivi du dossier médical ; il leur incombe de récupérer les images et/ou le compte rendu anatomo-pathologique des lésions. Ces informations sont récupérées par courrier/fax via les cabinets d'anatomo-pathologie ou directement depuis les patients. Ces données sont ensuite ressaisies dans le logiciel métier des associations par le médecin coordonnateur.

Problèmes soulevés

Le problème est simple : Il n'existe aucun moyen automatique, ou semi-automatique d'assurer le transport de l'information depuis les laboratoires d'anatomo-pathologie vers les associations chargées du dépistage organisé.

3.1.2 Pour les cabinets d'anatomo-pathologie

Les cabinets d'anatomo-pathologie n'ont à l'heure actuelle aucun export de données qui est fait, ces comptes-rendus anatomo-pathologiques restent le plus souvent à l'intérieur des bases de données de leur logiciel de gestion. A la demande des associations, elles leur expédient manuellement des données.

3.1.3 Pour la santé publique

La santé publique dispose de registres nationaux pour l'établissement ou l'évaluation de critères sur la santé publique (incidence, prévalence etc.). A l'heure actuelle ils ne disposent pas des données anatomo-pathologiques pour appuyer leurs études statistiques.

3.1.4 Contraintes

Les contraintes sont aussi fortes : il s'agit de données médicales nominatives et la loi française sur la protection des individus face au traitement informatique des données est très restrictive et règlementée.

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 3 : ANALYSE	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	14 sur 38

Par ailleurs, les propriétaires des données anatomo-pathologiques ne souhaitent pas voir leurs données exportées sans leur accord dans un autre lieu que leurs locaux. Ils possèdent des conventions avec les associations de dépistage pour des traitements de données au cas par cas. Ils souhaitent ainsi rester maîtres de leurs données, en étant les uniques propriétaires. Une exportation massive et complète de ces données entraînerait un refus catégorique de leur part.

3.2 Analyse de l'existant

3.2.1 Logiciels Métier

ARDOC/ABIDEC : Zeus – Editeur OSI-SANTE

Les associations de dépistage organisé ont à leur disposition un logiciel fondé sur le/la bénéficiaire placé au centre du dispositif.

Il utilise les technologies du WEB et Internet avec renforcement de partenariats tels que ANTARES - Acxiom (cartographie) - Fenics dans le domaine de l'imagerie médicale numérique...

Langage de développement : Visual Studio.net ; VB.Net et Asp.Net

Base de données : Microsoft SQL Server 2000 ou 2005

Anatomo-pathologistes : DIAMIC – Editeur INFOLOGIC

Les cabinets d'anatomo-pathologistes de la région utilisent le logiciel DIAMIC qui permet d'effectuer l'ensemble de gestion informatique de dossiers anatomo-pathologiques.

Requêteur Diamic

Le logiciel est fourni avec un environnement permettant d'exécuter des requêtes directement sur la base de données interne. De plus, il est possible de l'interroger directement via le langage SQL.

Santé publique

Le service de Santé Publique utilise des logiciels de traitement des données (statistiques et épidémiologiques tels qu'Epi Info, Epi Data, SAS, MEDCALC, STATA, SPSS). Ces logiciels peuvent fonctionner avec des données texte, EXCEL, DBASE, ACCESS ou selon le format propre à chaque logiciel de traitement utilisé précité.

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 3 : ANALYSE	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	15 sur 38

3.3 Objectif

3.3.1 Dépistage par les associations

But : Suivre ce qui se passe après un dépistage positif (recherche de la VPP*, du parcours de soins post dépistage positif,...) ou négatif (recherche de la VPN* et/ou des cancers de l'intervalle,...) pour le patient. Le recueil est de type prospectif (les données sont collectées après l'invitation/la réalisation du dépistage) de façon itérative et/ou au fur et à mesure de la survenue des évènements.

Les associations sont tenues d'assurer le suivi médical des personnes positives. Pour cela, le recueil des explorations complémentaires est nécessaire pour obtenir un diagnostic et déterminer le caractère bénin ou malin de la tumeur. L'institut National de Veille Sanitaire évalue et valide annuellement les données des programmes de dépistage organisé.

Utilisateurs : les associations de dépistage

Le schéma ci-dessous montre l'objectif principal pour les associations de dépistage, il s'agit de créer un réseau de communication entre les bases de données anatomo-pathologiques et les bases propres aux associations pour transférer :

1. les rapports anatomo-pathologiques des personnes suivies dans le cadre du dépistage organisé, mission première des associations ;
2. obtenir des données épidémiologiques sur les types de cancers dont les associations ont pour mission d'assurer le dépistage organisé : le cancer du sein et du colon.

A terme, tous les cancers devront pouvoir être repérés.

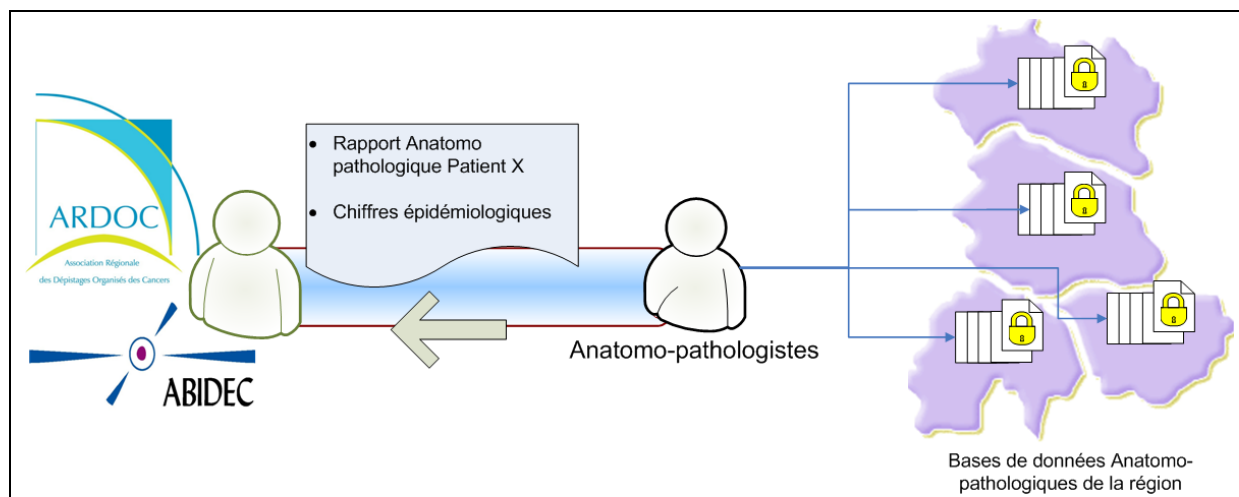


Figure 1 - Objectif Associations

-
- * VPP = valeur prédictive positive
 - * VPN = valeur prédictive négative

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 3 : ANALYSE	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	16 sur 38

Veille sanitaire

But : Obtenir des infos de type veille ou observation sanitaire sur l'incidence (nombre de nouveaux cas de cancers), la prévalence (nombre total de malades), l'évolution dans le temps, l'évolution des stades au dépistage, recherche éventuelle de cas groupés (clusters)...

Utilisateurs : Santé publique : CHU, DRASS (inspection régionale de la santé), CIRE, avec parfois une demande pressante d'une autorité administrative (préfet/maire,...), potentiellement l'InVs ou éventuellement la DGS¹.

L'identification du patient ne se pose pas ; il suffit d'avoir des données sans doublons mais anonymes. L'interrogation des bases de données pour mettre à jour les statistiques doit être périodique, à définir entre hebdomadaire et journalière.

Exemple concret : une autorité régionale (ex : préfet) peut être sollicitée sur le nombre de cancers auprès d'une usine de traitement des déchets ; ce réseau devrait dans un temps relativement court permettre de répondre à cette question (la difficulté sera de vérifier la pertinence des résultats).

Le schéma ci-dessous montre l'objectif principal de la veille sanitaire, à savoir récupérer des informations statistiques disponibles sur l'ensemble de la cancérologie dans les bases de données anatomo-pathologiques.

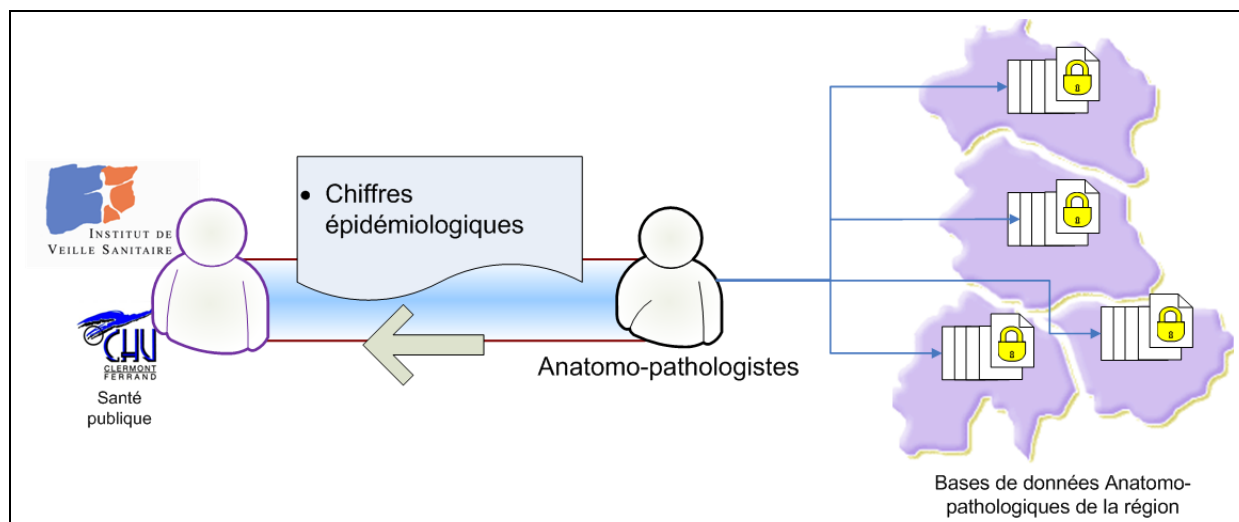


Figure 2 - Objectif épidémiologie

¹ Direction Générale de la Santé

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 3 : ANALYSE	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	17 sur 38

3.3.2 Indicateurs évaluatifs

But : Les programmes de dépistage organisé des cancers comportent des indicateurs pour évaluer leurs efficacités. Il s'agit d'obtenir via le réseau sentinelle les données nécessaires au calcul de ces indicateurs de la façon la plus efficiente possible (type d'examens cytologiques et résultats histologiques).

Utilisateurs : Associations chargées du dépistage organisé (ARDOC/ABIDEC), structures de santé publique à tous niveaux.

Les indicateurs statistiques d'efficacité pour les associations sont les suivants :

- Taux de cancers dépistés
- Taux de cancers intra-canalaires stricts (CICS)
- Taux de cancers invasifs de taille inférieure ou égale à 10 mm
- Taux de cancers micro-invasifs ou invasifs sans signe d'envahissement ganglionnaire

Le calcul de ces indicateurs est possible par la transmission par les anatomopathologistes des résultats des examens cyto-histologiques de diagnostic (biopsies chirurgicales, cytoponction, micro-biopsie, macro-biopsie et diagnostic histologique pré-opératoire).

3.3.3 Scientifique

Ce projet, de part l'implication de nombreux acteurs scientifiques devrait être le support de publications scientifiques de plusieurs ordres :

D'un point de vue médical il serait possible d'obtenir des études épidémiologiques ou statistiques inédites sur la cancérologie.

D'un point de vue informatique la solution technique pourrait proposer des innovations sujettes à des publications dans le domaine des technologies de l'information et des communications.

Ces publications scientifiques devront mentionner le réseau sentinelle comme source de l'obtention des données exploitées.

3.4 Perspectives (objectif)

Ce projet ne doit pas être voué à devenir une application monolithique dédiée à l'échange de données entre, d'un côté les acteurs du dépistage organisé et la santé publique et de l'autre les anatomo-pathologistes. Il doit être conçu de façon à pouvoir être étendu à d'autres problématiques similaires liées à l'échange de données médicales entre différents acteurs avec une forte contrainte de sécurité et de respect de la vie privée.

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 3 : ANALYSE	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	18 sur 38

L'architecture technique devra être conçue avec une certaine souplesse et flexibilité, qui permettra un ajout d'acteurs de façon progressive. Cette conception devra répondre aux exigences du cahier des charges en termes de perspectives d'évolution.

A l'heure actuelle, la plupart des solutions proposées pour résoudre ce type de demande consistent à extraire de façon périodique les bases de données concernées pour les rassembler en un endroit unique. C'est par exemple la solution proposée par l'InVS pour la surveillance épidémiologique des cancers fondée sur des registres nationaux.

Ces solutions comportent de nombreux inconvénients :

- le coût élevé de la collecte des données, de la vérification de leur pertinence et la maintenance des systèmes centraux. (4,2Millions d'€ en 2008) ;
- la couverture limitée de la population, avec une date de disponibilité des données fiables souvent dans un délai de 4 ans (En 2008, mise à disposition des données 2004).

Par ailleurs l'initiative nationale de création du DMP (Dossier Médical Patient) a échoué pour des raisons techniques et aussi budgétaires. A l'origine ce projet visait à rassembler à un même endroit les données médicales de tout patient sans pour autant avoir de lien entre les informations.

Depuis, de nouvelles initiatives ont vu le jour, avec l'apparition d'initiatives comme le GIP-DMP¹ et ses 14 projets régionaux de constitution de ce dossier (Projet SIMPA pour l'Auvergne ou DPPR pour la région Rhône-Alpes). L'initiative du DPPR consiste à mettre en réseau les différentes sources d'informations pour, une fois l'ensemble des sources rassemblées, pouvoir constituer ce dossier médical.

3.4.1 A court terme

La perspective de ce projet est d'adapter les nouvelles technologies de l'information et de la communication et plus particulièrement l'essor récent en ressources réseau qui, plutôt que de centraliser des bases de données, les rendent disponibles au travers des réseaux, et plus particulièrement Internet.

De la même manière il serait intéressant d'appliquer ces mêmes technologies à l'échange d'images médicales nécessaires aux associations pour le dépistage du cancer du sein par exemple.

3.4.2 A plus long terme

L'intégration de nouveaux acteurs pourrait commencer par la surveillance des mélanomes, ce cancer faisant déjà l'objet d'un recueil de données cliniques (dermatologues) et anatomo-pathologique dans la région Auvergne depuis plus de 10 ans. L'OBRESA assure le suivi, l'exploitation et la publication des résultats.

¹ www.d-m-p.org

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 3 : ANALYSE	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	19 sur 38

A plus long terme de nouveaux acteurs, fournisseurs ou utilisateurs pourront intégrer le réseau sentinelle.

Nous pouvons citer, à titre informatif :

- Le DMP régional (SIMPA)

Dans une perspective d'élargissement on pourrait inclure :

- les caisses d'assurance maladie (par extrapolation au cancer de l'utérus par exemple)
- Le service d'oncogénétique du Pr. Yves-Jean Bignon (par extrapolation à la génétique)
- L'InVS, la DGS, l'INCa¹ et la FNORS².

3.5 Moyens du projet

Ils seront assurés dans la phase de réalisation du projet par une association loi 1901 selon des modalités qui seront définies ultérieurement.

3.6 Bénéfices attendus

3.6.1 Pour les associations :

Si ces associations disposaient d'un outil informatique pour récupérer ces données, les bénéfices réalisés seraient :

- Un recueil plus exhaustif des données
- Une meilleure structuration de l'information grâce aux comptes rendus standardisés
- Une limitation de risques d'erreur de recopie / d'interprétation
- Un gain de temps de saisie, en affranchissement et en papier

Globalement, elles pourraient augmenter leur réactivité et leur efficacité, qui sont des critères d'évaluation, mais aussi assurer de façon beaucoup plus précise et juste le suivi des personnes positives.

3.6.2 Pour les structures de santé publique

Pour la santé publique les bénéfices attendus sont essentiellement sur la possibilité nouvelle de requêtes statistiques sur l'ensemble des bases anatomo-pathologiques de la région, avec une exhaustivité théoriquement complète sur ces bases.

¹ Institut National du Cancer

² Fédération Nationale des Observatoires Régionaux de Santé

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 3 : ANALYSE	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	20 sur 38

Bénéfices :

- Disposer des données d'évaluation pour des requêtes prédéfinies à brève échéance : de quelques minutes à quelques jours suivant le type de requête
- Fonder des avis techniques sur des données de qualité
- Posséder des données fiables et vérifiées en matière d'épidémiologie
- Avoir un temps de réponse court, adapté avec des informations pertinentes et de qualité en cas de crise sanitaire, potentielle ou avérée, locale ou généralisée, d'où la nécessité d'une partie validation. [voir 8 : Validation p.34]

A plus long terme il serait possible de couvrir d'autres types de cancers, et d'inclure d'autres structures de santé publique. Le bénéfice serait une santé publique à plus grande échelle pour d'autres types d'études.

De la même manière il serait possible de faire un outil d'analyse de pandémie en temps quasi réel, en fonction des données accessibles à un instant et de la complexité de la requête.

Ces aspects seront pris en compte dans une phase ultérieure du projet.

3.6.3 Pour les cabinets d'anatomo-pathologie

Même s'ils devraient uniquement fournir des données, donc une charge supplémentaire de mise à disposition depuis leurs locaux, les bénéfices seraient visibles sur le temps passé par le secrétariat à renseigner les comptes-rendus anatomo-pathologiques aux associations, donc un gain logistique ou en télécommunication de ces dossiers.

Ils gagneraient en souplesse et en confort avec un échange automatique avec les associations.

Par ailleurs il sera possible pour les cabinets de s'appuyer sur des chiffres plus exhaustifs pour des publications scientifiques qu'ils produisent régulièrement. Ces études nécessitent des échanges avec d'autres partenaires, par exemple savoir la mortalité d'un type de cancer.

3.7 Liste des objectifs :

3.7.1 Fonctionnels

Le réseau devra :

- offrir une méthode de mise à disposition des données provenant des fournisseurs aux différents clients du projet
- offrir une méthode de requête à distance permettant l'intégration facile sous un format standardisé et normalisé dans des logiciels tiers appartenant aux clients du réseau

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 3 : ANALYSE	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	21 sur 38

- respecter la propriété des données en ne faisant pas d'extraction complète, de copie ou de déplacement de données en dehors des locaux des producteurs. Ainsi les données seront seulement consultées depuis les locaux des propriétaires

3.7.2 Techniques

Déploiement d'un réseau sentinelle :

- fortement sécurisé
 - identification forte par carte CPS, certificat
 - champ d'action encadré et protégé par l'utilisation de services web
- accessible à un utilisateur familier en informatique avec un matériel classique
- hébergement de la plateforme en lieu sûr et neutre (partie sécurité)

Fournir une méthode d'identification des patients :

- qui offrira de bonnes performances pour le recoupement de dossiers patients
- qui, en aucun cas, associera deux patients distincts à une même identité
- qui respectera les critères de la CNIL¹ et autres disposition juridiques

3.7.3 Administration et maintenance

Permettre aux administrateurs :

- de gérer des accès et des droits des utilisateurs
- de récupérer des traces d'utilisation des utilisateurs (à la demande)
- d'offrir un support au niveau utilisateur
- de suivre et maintenir les évolutions des différents modules applicatifs

3.7.4 Juridiques

Le projet devra :

- satisfaire les contraintes fixées par la CNIL
 - protéger les personnes du traitement informatique de leurs données
- mutualiser les différents acteurs (fournisseurs et clients des données) par le biais de documents contractuels, conventions et engagements
- Identifier fortement un utilisateur du réseau
 - garantir une traçabilité des accès au réseau et des échanges effectués

¹ Commission Nationale Informatique et Libertés

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 3 : ANALYSE	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	22 sur 38

3.8 Perspectives de solution

La mise en place d'un système d'information pour l'échange direct de données patient entre les cabinets d'anatomo-pathologie et les structures départementales en charge du dépistage organisé des cancers devrait permettre de réduire considérablement le coût et le risque d'erreur liés à la ressaisie des données patients. De plus, les acteurs de la surveillance épidémiologique des cancers, pourront à terme récupérer des données statistiques en interrogeant les bases de données (ARDOC/ABIDEC) ou directement celles des cabinets d'anatomo-pathologie ayant donné leur accord.

La solution proposée pourra s'appuyer sur le savoir-faire de l'équipe PCSV du laboratoire de Physique Corpusculaire dans le domaine des grilles de calcul pour l'adapter à la problématique présentée ici. En effet, les technologies liées aux grilles de calcul peuvent s'adapter à d'autres applications que le calcul distribué intensif. Ce projet explorera en premier lieu les solutions offertes par cette technologie pour mener à terme ce projet.

Pour la partie expérimentale du réseau sentinelle, il sera plus commode et pratique de commencer par la gestion du cancer du sein entre un cabinet d'anatomo-pathologie ainsi qu'une association.

Pour la partie réalisation du projet, la société Maat-G propose de collaborer au projet en apportant son expertise dans le domaine technique pour la réalisation.

D'un côté juridique, il est possible de faire appel aux services du laboratoire de physique corpusculaire pour remplir les déclarations CNIL et au service de radiologie du Dr. Isnard pour l'aspect conventions entre acteurs et charte réseau.

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 4 : DESCRIPTION FONCTIONNELLE	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	23 sur 38

4 Description fonctionnelle

4.1 Caractéristiques et fonctionnement du système

La figure ci-après représente une formalisation des échanges entre les associations et les pathologistes réalisés par le réseau sentinelle.

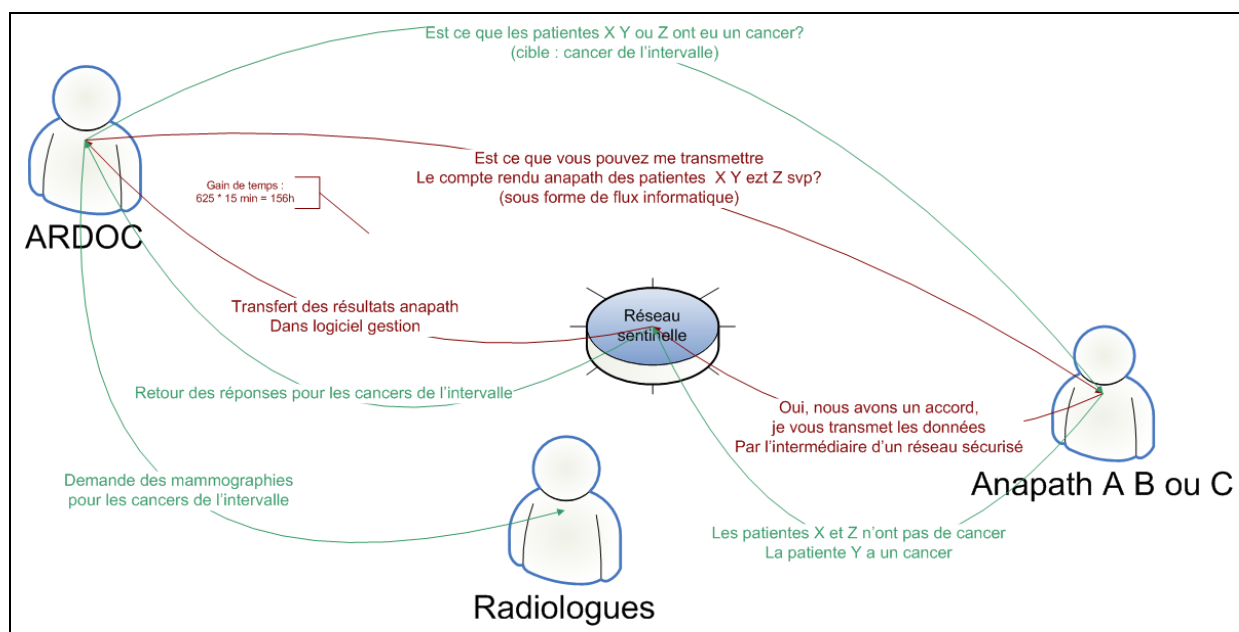


Figure 3 - Fonctionnement du système

4.1.1 Description des parties client

Les clients de données, comme les associations ou les structures de santé publique devront se charger d'intégrer, d'exploiter les données reçues du réseau sentinelle. Les messages seront normalisés selon une description standard des données pour faciliter leur intégration.

Les clients, s'ils souhaitent une intégration dans leurs logiciels métier devront le réaliser par eux-mêmes. Par ailleurs, en ce qui concerne les associations de dépistage organisé, l'éditeur OSI-Santé a été contacté pour réaliser cette intégration de données dans le logiciel Zeus.

Pour la santé publique, les données seront reçues sous forme de chiffres bruts suivant un standard qui sera défini dans un volet technique du projet. Une documentation sera fournie pour savoir comment interroger, récupérer et interpréter les résultats.

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 4 : DESCRIPTION FONCTIONNELLE	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	24 sur 38

4.2 Scénarios d'utilisation du système par acteur

Les parties suivantes représentent les différents cas d'utilisation par acteur du réseau sentinelle.

4.2.1 Scénario d'utilisation par les anatomo-pathologistes

Les laboratoires d'anatomo-pathologie sont les principaux fournisseurs des données et auront accès à une interface leur permettant d'extraire leurs données vers le serveur relié au réseau sentinelle qui leur est dédié. Ils pourront aussi interagir avec le serveur de sécurité pour définir les différents droits d'accès d'un utilisateur du système avec leurs données et consulter l'historique des accès à ses données par utilisateur.

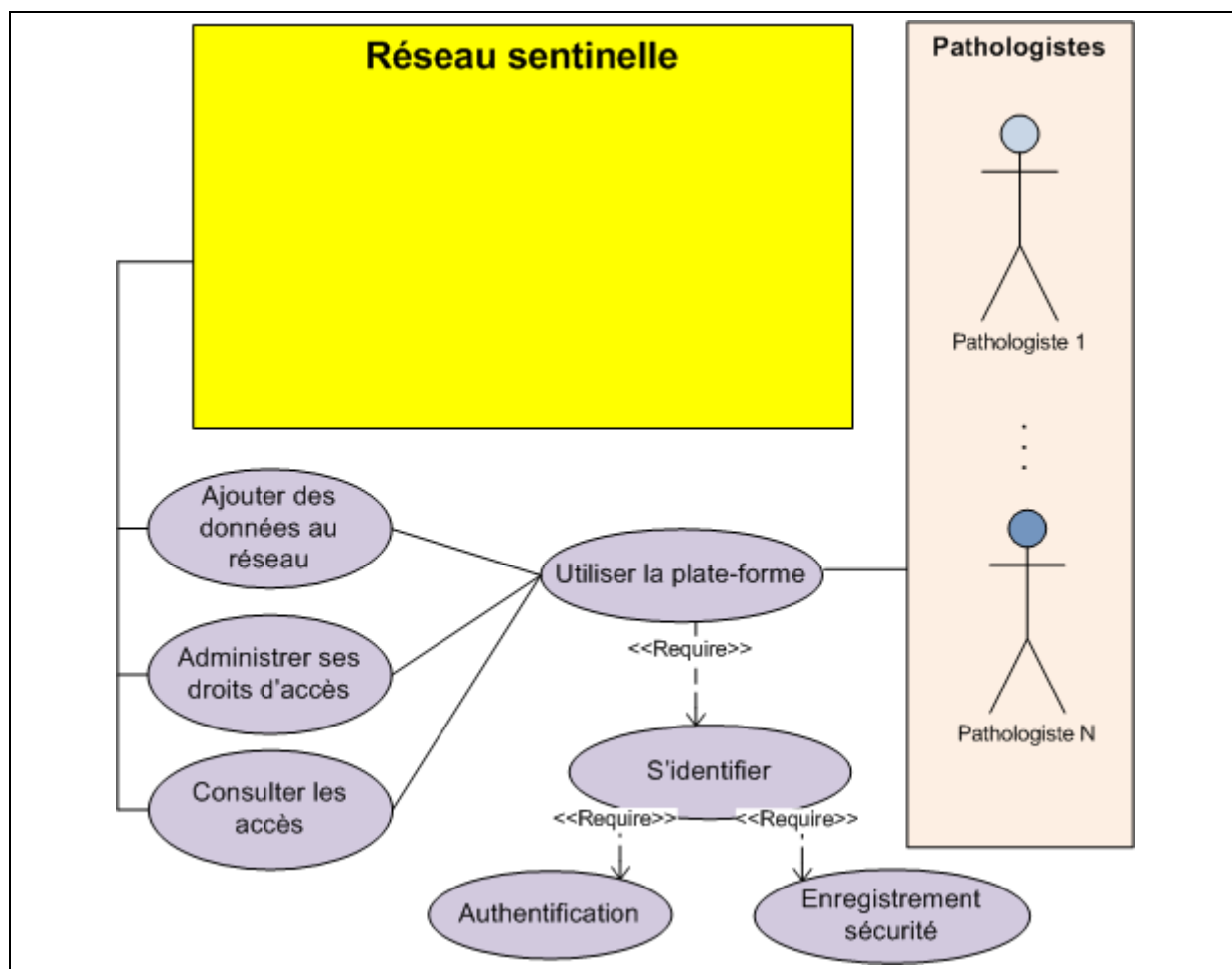


Figure 4 - Cas d'utilisation pour les pathologistes

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 4 : DESCRIPTION FONCTIONNELLE	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	25 sur 38

4.2.2 Scénario d'utilisation par les administrateurs

Les administrateurs auront à disposition des outils leur permettant d'ajouter des utilisateurs, de vérifier le fonctionnement du réseau, de consulter les rapports d'accès aux données etc.

Bien entendu, les administrateurs ne seront pas responsables de la gestion des droits d'accès des utilisateurs aux données des fournisseurs, ce sont ces producteurs de données et eux seuls qui fixent les droits et limites d'accès aux autres utilisateurs à leurs données. [Voir 6.1.2 : Gestion des habilitations p.31]

La figure ci-dessous résume différents scénarios d'utilisation du système pour les administrateurs :

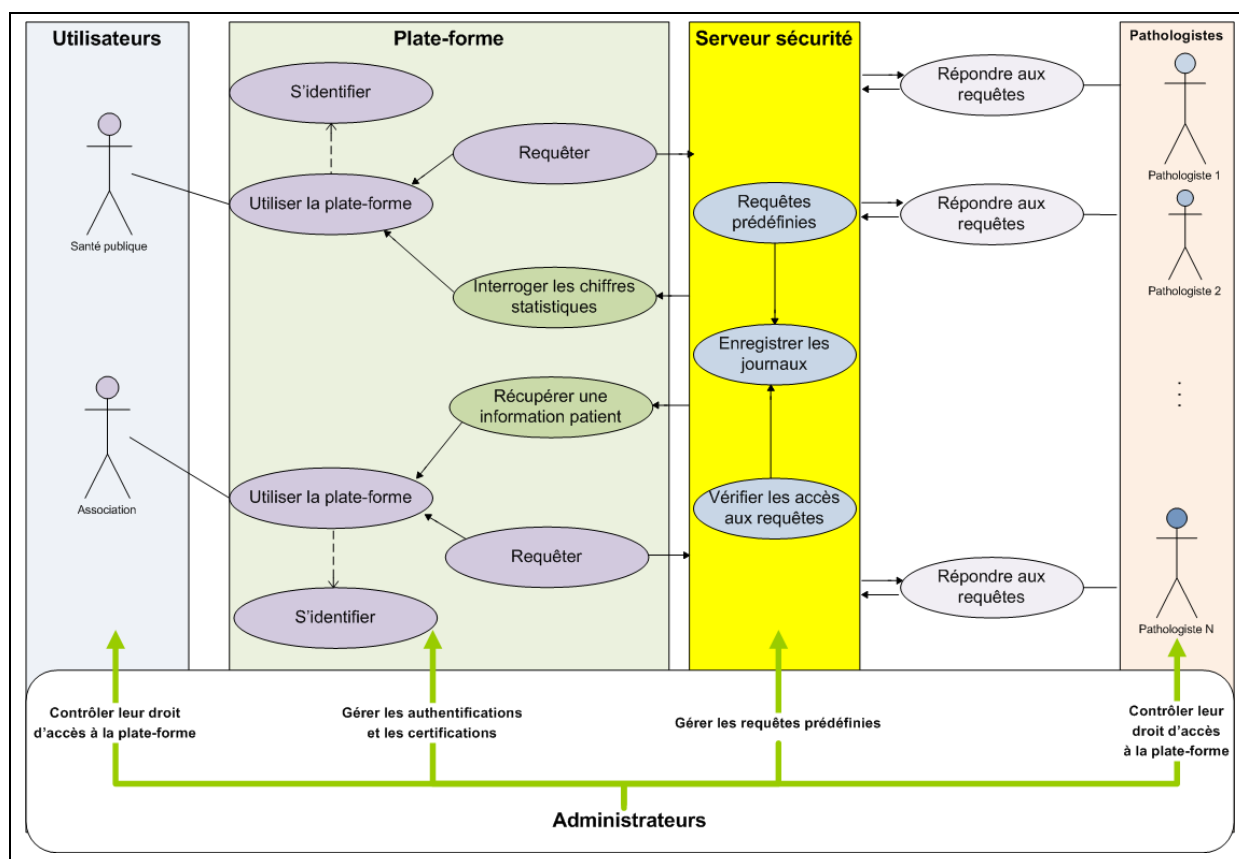


Figure 5 - Scénario d'utilisation Administrateurs

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 4 : DESCRIPTION FONCTIONNELLE	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	26 sur 38

4.2.3 Scénarios d'utilisation par les associations de dépistage organisé

La requête principale est la demande d'un compte rendu anatomo-pathologique au système. Le système interroge simultanément les différentes bases de données accessibles à l'association qui a effectué la demande (celles pour lesquelles les cabinets d'anatomo-pathologie ont donné leur accord). Une fois la réponse obtenue, il structure les informations et les renvoie à l'utilisateur. Cette opération pourra alors être effectuée pour chaque patient en quelques secondes (dépendra de la disponibilité du réseau sentinelle). Le logiciel métier sera capable d'interpréter les résultats et de les insérer directement dans la base de données locale, sous réserve d'un développement spécifique de la part de l'éditeur.

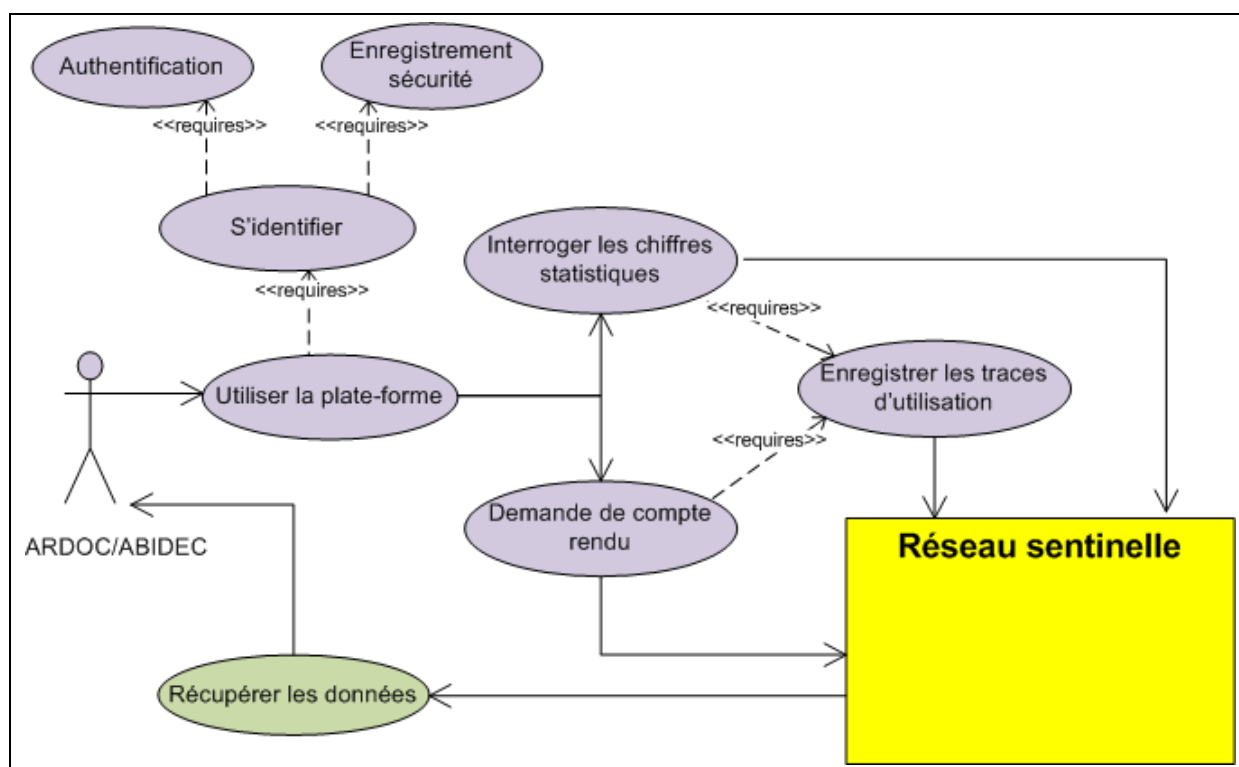


Figure 6 - Scénario d'utilisation pour les associations

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 4 : DESCRIPTION FONCTIONNELLE	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	27 sur 38

4.2.4 Scénarios d'utilisation par les structures de santé publique

Les structures de santé publique auront besoin d'évaluer le nombre de cancer du sein traité chaque année, des indicateurs de performance du dépistage organisé, de l'évolution du nombre et du type des cancers, éventuellement de données infrarégionales pour le repérage de cas anormalement élevés dans un territoire donné. Pour cela elles adresseront leur demande au système qui interrogera de la même façon les différentes bases de données anatomo-pathologiques auxquelles la structure de santé publique concernée aura accès ou encore les bases de données des associations inscrites en tant que fournisseurs de données (2^{ème} phase). Le système sera capable de produire différentes statistiques/études épidémiologiques sur l'incidence des cancers en Auvergne.

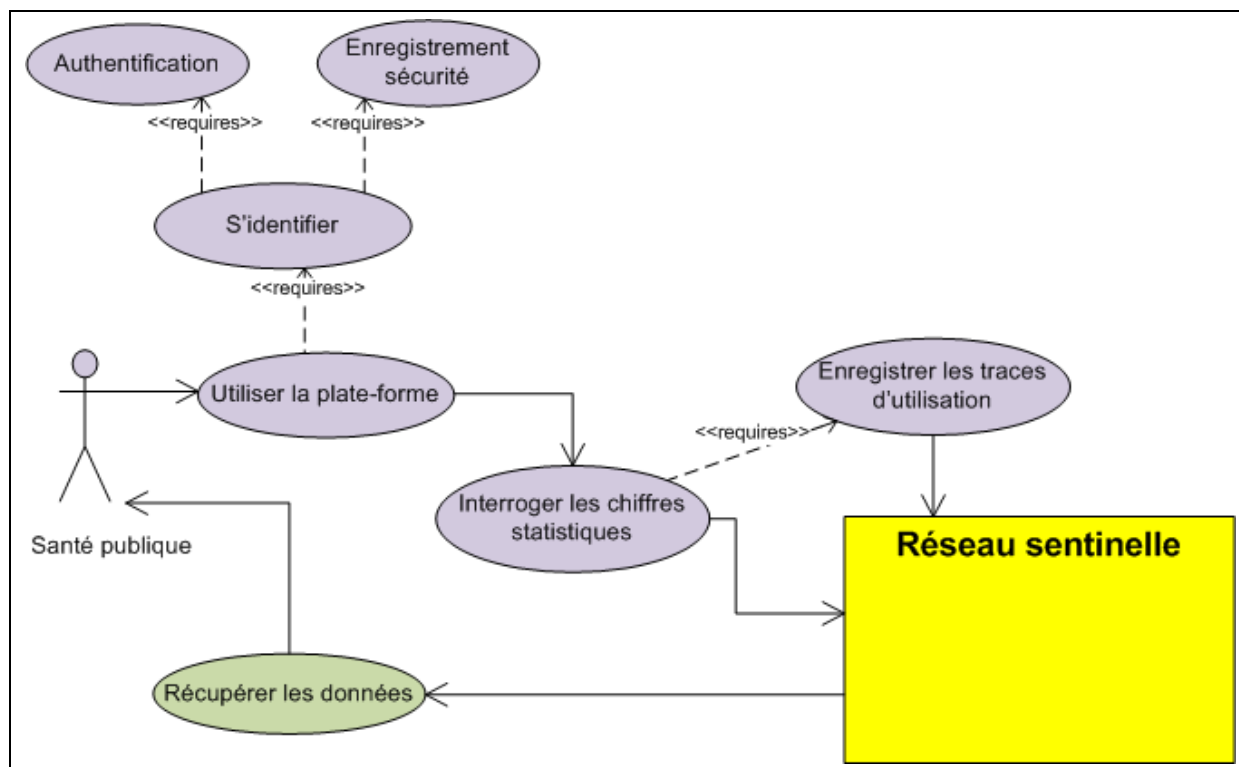


Figure 7 - Scénario d'utilisation pour les acteurs de santé publique

4.2.5 Imagerie médicale

L'accès aux images radiologiques devra être pris en compte lors de la conception du projet, même si la mise en œuvre se fait dans un second temps. Une des solutions à étudier pour la réalisation du transfert des images sera MDM¹, qui permet l'accès par la grille aux images issues directement d'un serveur DICOM.

¹ Medical Data Management : Accès aux serveurs d'images médicales depuis la grille.

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 5 : DONNEES	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	28 sur 38

5 Données

5.1 Types de données

5.1.1 Données patients

Les données anatomo-pathologiques seront décrites dans l'édition d'un compte rendu standardisé appelé bible des données. [Voir Bible des données anatomo-pathologiques standardisée p.35]

5.1.2 ADICAP

Association pour le Développement de l'Informatique en Cytologie et Anatomo-Pathologie

Le compte rendu anatomo-pathologique comporte impérativement un ou plusieurs codes ADICAP, qui est un code français pour chaque diagnostic anatomo-cyto-pathologique, pour chaque lésion, cancéreuse en particulier ainsi que leur gravité. Ce code est techniquement 'interchangeable' avec le code OMS (mondial cette fois).

Il peut exister plusieurs codes ADICAP par pièce opératoire.

Les associations de dépistage organisé se servent du code ADICAP.

5.1.3 Images médicales (Mammographies...)

Dans un second temps, le réseau devrait pouvoir récupérer des images médicales depuis les cabinets de radiologie. La forme et les modalités seront précisées ultérieurement.

5.2 Données par Acteur

5.2.1 Associations de dépistage organisé

Les associations ont besoin périodiquement d'informations statistiques pour mieux juger l'impact du dépistage organisé et leur performance.

Les associations ont besoin, pour chaque personne invitée au dépistage et positive de récupérer le compte rendu anatomo-pathologique standardisé pour assurer le suivi de cette personne.

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 5 : DONNEES	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	29 sur 38

5.2.2 Pathologistes (descriptifs des fiches de comptes-rendus standardisés)

Les pathologistes pourront fournir un compte rendu standardisé pour faciliter l'échange de données sur le réseau. Cette normalisation facilitera aussi bien le fonctionnement des cabinets que celui des associations chargées de relire ces comptes rendus. [Voir Bible des données anatomo-pathologiques standardisée p.35]

Ces données sont à récupérer depuis les bases de données des anatomo-pathologistes.

5.2.3 Statistiques/Epidémiologie

La figure suivante présente les données ou indicateurs nécessaires à l'export pour l'épidémiologie dans le cas particulier du cancer du sein:

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 5 : DONNEES	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	30 sur 38

Pathologistes	Associations	Cabinet de Radiologie	Santé publique	Données brutes nécessaires	Indicateurs épidémiologiques
✓				Laboratoire concerné	Exhaustivité régionale
✓				Sexe (inclut les hommes, soit 1% des cancers)	Fréquence par sexe
✓	✓			Prescripteur	Dépistage individuel ou collectif Flux intra ou interrégional
✓				Age/Date de naissance	Moyenne d'âge des cancers détectés
✓				Commune et département de domicile	Flux intra ou interrégional Variable de comparaisons par gravité
✓				Critères d'identification des bénéficiaires (N° SS, nom patronymique, adresse...)	Identification pour élimination de doublons
✓				Date de réception au laboratoire Date d'examen	Délais avant diagnostic
		✓		Date de la mammographie	Délais avant diagnostic
✓	✓			Antécédent de cancer du sein	Pourcentage de récurrence
✓				Antécédent d'autre cancer (code ADICAP)	
✓				Cytologie/histologie/biopsie chirurgicale Ou cytologie/histologie	Type d'analyse pratiquée
✓				Taille	Gravité de l'atteinte
✓				Type : in situ, invasif...	Type de cancer
✓				Atteinte ganglion	Gravité de l'atteinte
✓				Classifications (TNM...)	Gravité de l'atteinte, INCIDENCE des cancers du sein
✓				Autres facteurs pronostics	
✓				Facteurs prédictifs	

Figure 8 - Données nécessaires à l'épidémiologie

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 6 : SECURITE	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	31 sur 38

6 Sécurité

6.1 *Identification utilisateur*

6.1.1 Carte CPS

La carte CPS est une carte électronique individuelle protégée par un code confidentiel, aujourd'hui diffusée à 600 000 exemplaires auprès des personnels de santé (professionnels de santé et personnels auxiliaires). Elle contient des informations portant sur l'identité du professionnel de santé, sa qualification, ses différentes situations d'exercice. Elle contient aussi des données de facturation pour l'établissement des feuilles de soins électroniques, dans le cadre de l'application SESAM Vitale (plus de 900 millions de feuilles de soins électroniques réalisées en 2006).

La carte CPS contient en outre les « certificats CPS », qui constituent de véritables pièces d'identité dématérialisées et certifiées par le GIP-CPS. Ils sont utilisés par leur titulaire comme gages de confiance, dans le cadre des applications communicantes, mettant en œuvre des données de santé confidentielles, et permettent l'identification du professionnel de santé, son authentification, une signature électronique et un chiffrement des données.

Son utilisation serait plus que souhaitable pour la mise en place d'un tel réseau, vu que les utilisateurs seraient tous issus ou en relation avec le monde médical.

6.1.2 Gestion des habilitations

Afin d'assurer le respect de la propriété des données, condition impérative pour que les anatomo-pathologistes acceptent un transfert de leurs informations, une interface dédiée à l'administration des droits par ces acteurs devra être implémentée pour garantir cette propriété.

Cette interface permettra de couper l'accès à leur serveur de données du réseau sentinelle en fonction des utilisateurs. Ainsi, le fournisseur des données aura la certitude de l'utilisation qui est faite de ses données. L'administrateur n'a pas l'accès à cette partie, qui doit être maintenue par le propriétaire des données.

La figure suivante représente un exemple d'interface de gestion des droits d'accès pour un fournisseur de données.

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 6 : SECURITE	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	32 sur 38

Clients Fournisseurs	ARDOC	ABIDEC	Santé Publique
<input type="checkbox"/> Anatomo-Pathologiste A			
<input type="checkbox"/> Anatomo-Pathologiste B			
<input type="checkbox"/> Cabinet A			
<input type="checkbox"/> Centre de lutte contre le cancer			
<input type="checkbox"/> Service d'Anatomo-Pathologie A			
<input type="checkbox"/> Service d'Anatomo-Pathologie B			
<input type="checkbox"/> Service d'Anatomo-Pathologie C			

Figure 9 - Exemple d'interface de gestion des droits

6.2 Identification patient

L'identification du patient est un point crucial pour mener à bien ce projet. Bien que cette information ne soit pas nécessaire pour les études épidémiologiques où un « simple » dédoublonnage suffit, cette information est primordiale pour l'export des données nominatives issues des rapports anatomo-pathologiques vers les associations.

Le but de l'identification patient est de rapprocher de part et d'autre du réseau un identifiant à un autre, en se servant des informations disponibles sur le patient : l'état civil, l'adresse, etc.

Dans le cadre de Lifegrid¹ est né un projet nommé « Système d'information pour la généralisation de l'accès sécurisé à la base de données régionale d'hémovigilance ». Ce projet vise à assurer la traçabilité des poches de sang. Pour cela il a été nécessaire de centraliser dans une même base de données les différents identifiants propres aux structures de santé où est enregistré un patient. Ainsi, en couplant identifiant de la structure de santé (hôpital, clinique, médecin etc.) et le numéro interne du patient il est possible d'avoir un ensemble de plusieurs identifiants pour un patient. De cette manière il serait possible d'ajouter l'identifiant interne aux associations ou aux cabinets d'anatomo-pathologie dans cette base et pouvoir ainsi vérifier la concordance de l'information patient. Ce projet est animé par le Docteur Anne DOLY du Centre Jean Perrin.

¹ <http://www.lifegrid.fr/>

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 7 : ECHEANCIERS (MISE EN ŒUVRE)	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	33 sur 38

7 Echéanciers (mise en œuvre)

Dates	Tâche
Juillet 2008	Plan du cahier des charges
	Réunion technique 1 :
	Scénarios d'utilisation :
	(Association/Anapath/Epidémiologie)
28 Juillet 2008	Sécurité (CNIL, CPS, Identification patient)
	Images
	Echéancier
Fin Septembre 2008	Cahier des charges, version préliminaire
	Réunion technique Déclarations CNIL
Début Octobre 2008	Discussion du cahier des charges
	Discussion échéancier/plan de financement
	Conventions
Novembre 2008	Cahier des charges, première version
	Réunion de travail :
Fin Novembre 2008	Discussion du cahier des charges
	Plan de financement (livraison)
	Avancement déclaration & conventions
Fin 2008	Cahier des charges version finale (livraison)
	Préparation du cahier technique
2009 ?	Implémentation

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 8 : VALIDATION	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	34 sur 38

8 Validation

La partie concernant la validation des résultats statistiques renvoyés par le réseau sentinelle est un volet très important pour la suite du projet. Le service de santé publique du CHU de Clermont-Ferrand propose son savoir faire dans le domaine pour mener à bien ce travail.

Les spécifications et objectifs de cette partie seront définis dans un document séparé ultérieurement.

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
ANNEXES	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	35 sur 38

9 ANNEXES

9.1 *Bible des données anatomo-pathologiques standardisée*

Bible des données anapath

CODE : NOM : ADICAP :

Tumeur mammaire

Renseignements cliniques :

Macroscopie :

Tumorectomie, Mastectomie : ____

Coté : ____

Quadrant : ____

Volume pièce : ____x____x____cm

Lambeau cutané : ____

Mamelon : ____

Poids : ____ g

Taille de la lésion : ____

Recoupe(s) : ____

Repérage par fil métallique : ____

Ganglion sentinelle : ____

Nombre : ____

Curage axillaire : ____

Autre : ____

Histo Pathologie

Tumeur :

Examen extemporané effectué : ____ sur tumeur : ____ limites : ____ ganglions : ____ Résultat : ____ confirmé

Type histo-pathologique : carcinome ____ infiltrant

Carcinome in situ associé : ____% Type : ____ de grade : ____ Nécrose type comédocarcinome : ____

Cicatrice de prélèvement antérieur : ____

Mamelon atteint : ____ Si oui par : Paget CCI CIS Emboles intra vasculaires dermiques

Emboles vasculaires périphériques : ____

Calcifications retrouvées sur lames : ____

Multifocalité : ____

Taille tumorale définitive : ____mm

Grade histo-pronostique de Scarff, Bloom et Richardson modifié par Elston et Ellis (Nottingham) : ____

Différenciation : ____ Anisonucléose : ____ Mitoses : ____

Recoupe(s) : ____

Exérèse complète : ____

Plus petite distance séparant le tumeur (composante infiltrante, in situ, embole) de la marge la plus proche > : ____ mm

Autres foyers : ____ Taille : ____ Siège : ____

Autre : ____

Ganglion(s) lymphatique(s) :

Ganglion(s) sentinelle(s) : ____ dont métastatiques : ____ micrométastase : ____ rupture capsulaire : ____

Ganglion(s) autres ou curage standard : ____ dont métastatiques : ____ rupture capsulaire : ____

Immuno Histo Chimie ____

Conclusion

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 9 : ANNEXES	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	36 sur 38

9.2 Glossaire (vocabulaire)

Vocabulaire informatique

Firewall : Un firewall est un élément du réseau informatique, logiciel et/ou matériel, qui a pour fonction de faire respecter la politique de sécurité du réseau, celle-ci définissant quels sont les types de communication autorisés ou interdits.

Vocabulaire grille de calcul :

Grille informatique : C'est une architecture informatique permettant de distribuer des calculs ou du stockage sur de nombreuses machines en réseau. Contrairement à un cluster, une grille peut être de taille mondiale.

MDM : Medical Data Management : Mécanisme passerelle entre les technologies liées à la grille et les serveurs DICOM.

Vocabulaire médical :

DICOM : Digital Imaging and COmmunications in Medecine est un standard de communication et d'archivage en imagerie médicale. C'est aussi par extension le format de fichier faisant référence dans le domaine de l'imagerie médicale. Il permet de sauvegarder des informations administratives du patient, des données d'acquisition de l'image, ainsi que toutes les caractéristiques permettant de traiter l'image sur une console DICOM compatible.

PACS : Picture Archiving and Communications Systems sont des machines ou des réseaux dédiés au stockage, à l'accès, à la distribution et à la présentation d'images médicales.

Mammographie : La mammographie a pour but de déceler au plus tôt des anomalies avant même qu'elles n'aient provoqué des symptômes cliniques. Elle peut permettre, ainsi, de détecter des cancers bien avant qu'ils ne soient palpables.

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
ANNEXES	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	37 sur 38

10 Crédits

10.1 Acteurs techniques

Pr. Jean-Yves Boire

ERIM - Faculté de médecine, service de santé publique, CHU

j-yves.boire@u-clermont1.fr

Dr. Vincent Breton

Directeur de Recherche CNRS/IN2P3 - Laboratoire de Physique Corpusculaire

breton@clermont.in2p3.fr

Dr. Lydia Maigne

Maître de Conférences CNRS/IN2P3 - Laboratoire de Physique Corpusculaire

Convention avec le Centre Jean-Perrin (Denise Donnarieix)

maigne@clermont.in2p3.fr

Pierre Bouchet

Ingénieur informatique - Pole Santé République

pbouchet@polesanterepublique.fr

Paul De Vlieger

Doctorant Informatique - Equipe de Recherche en Imagerie Médicale

vlieger@clermont.in2p3.fr

Yannick Legré

Matt-G France

ylegre@maat-g.com

David Manset

Matt-G France

dmanset@matt-g.com

Dr. Anne Doly

CLCC Jean-Perrin – Service Informatique

10.2 Autres acteurs

M. Christian Celdran

Directeur régional - DRASS Auvergne

Dr. Pâquerette Lonchambon

DRASS Auvergne

Pr. Yves-Jean Bignon

CLCC Jean-Perrin – oncogénétique - Cancéropôle – CLARA

Yves-Jean.BIGNON@cjp.fr

10.3 Anatomo-pathologistes

Dr. Alain Gaillot

Sipath, chargé du projet ARDOC

agaillot@hotmail.fr

Pr. Pierre Déchelotte

Hôtel Dieu, CHU

pdechelotte@chu-clermontferrand.fr

Pr. Jean-Louis Kémény

Gabriel Montpied, CHU

Pr. Frédérique Penault-Llorca

CLCC Jean-Perrin

fpenault@cjp.fr

12 novembre 2008	Réseau sentinelle cancer Auvergne	Version	5.2
PARTIE 10 : CREDITS	CAHIER DES CHARGES	Auteur	Groupe de travail Réseau sentinelle
		Page	38 sur 38

10.4 Associations

Dr. André Lautier

Président de l'ARDOC

M. Labreure

Président de l'ABIDEC

Chantal Mestre

ARDOC

mestre@ardoc.org

10.5 Epidémiologie

Pr. Laurent Gerbaud

épidémiologie, Service de santé publique, CHU

lgerbaud@chu-clermontferrand.fr

Dr. Marie-Ange Grondin

AHU-épidémiologie, Service de santé publique, CHU

magrondin@chu-clermontferrand.fr

Dr. Lemlih Ouchchane

ERIM Service de Biostatistiques, Service santé
publique, CHU

lemlih.ouchchane@u-clermont1.fr

Candy Auclair

Ingénieur statisticien, unité d'épidémiologie,
économie de la santé et prévention

Service de santé publique, CHU

cauclair@chu-clermontferrand.fr

CONCLUSION

L'écriture de ce cahier des charges, fruit d'un long travail de discussion avec les différents acteurs du projet, a permis de spécifier les objectifs et contraintes du développement de la solution. Il présente aussi les différents flux de données et interactions logicielles qui seront à mettre en place pour assurer la communication entre ces parties.

Ce cahier pose une base quasi-exhaustive des attentes du projet poussée jusqu'à l'exploitation épidémiologique des données et d'une partie validation de la pertinence des résultats. Il propose aussi quelques pistes d'extension du réseau à d'autres acteurs, cas d'utilisation et applications médicales autres que le cancer.

Le cahier des charges met aussi en évidence les défis que représente la mise en œuvre du Réseau Sentinelle Cancer Auvergne. Les exigences des fournisseurs de données anatomopathologiques et les différentes contraintes, fonctionnelles ou de sécurité, imposent une réflexion approfondie sur les outils et méthodes à employer pour remplir les objectifs. Plus particulièrement, la création d'un réseau d'échange de données nominatives de santé sans centralisation avec des acteurs hors du cycle clinique (dépistage organisé) est une chose encore inédite tant les barrières légales complexifient la mise en œuvre.

Ainsi, la réponse à ce cahier des charges (cahier technique) implique une considération poussée de l'ensemble de ces éléments pour proposer une réponse techniquement, scientifiquement et légalement viable.

Chapitre 3. Mise en œuvre du Réseau Sentinelle Cancer Auvergne

INTRODUCTION

L'objectif de ce chapitre est de fournir une réponse technique au cahier des charges présenté au [Chapitre 2]. Il justifie les choix technologiques et architecturaux de ce qui deviendra le *Réseau Sentinelle Cancer Auvergne*.

Les éléments de ce chapitre présentent les choix techniques adoptés pour la conception de l'architecture de RSCA, la mise en place de l'architecture et son application dans le cadre du dépistage organisé des cancers en Auvergne.

- en premier lieu, les différents systèmes et standards de gestion de l'information médicale ont été étudiés dans le but de proposer une solution conforme au cahier des charges. Cette étape nécessite une comparaison des systèmes existants permettant la gestion et l'interconnexion de données. Une analyse brève des avantages et inconvénients respectifs de ces différentes méthodes permettra d'orienter les futurs choix du réseau ;
- ensuite vient la mise en place de l'architecture de RSCA, à partir des composants existants de grille présentés au préalable. Les choix de conception seront détaillés et les réalisations spécifiques nécessaires à la mise en application de l'architecture pour la problématique du dépistage organisé des cancers du sein et du colon seront présentées ;
- enfin, l'accès aux données depuis les structures d'anatomopathologie est détaillé ainsi que leur restitution dans les deux cas d'utilisation principaux du réseau, c'est-à-dire le transfert des comptes rendus médicaux vers les structures de dépistage d'une part et l'accès macroscopique à l'ensemble des données à destination de la santé publique d'autre part. Une attention particulière est apportée sur la structuration et la représentation générique des données ainsi que l'interfaçage avec les acteurs médicaux de RSCA.

Ce travail s'achève pour mettre en exergue les développements encore nécessaires à la réalisation de tous les objectifs du cahier des charges en respectant ses contraintes fonctionnelles.

3.1. LES SYSTEMES ET STANDARDS DE GESTION DE L'INFORMATION DE SANTE

3.1.1. *Les systèmes d'information et de communication de la santé*

Au travers des différents établissements médicaux, la gestion des l'information est une tâche vitale. Cependant chaque structure médicale a le plus souvent développé sa propre infrastructure de gestion de données adaptée à son environnement, le plus souvent unique à chaque établissement. Il existe de nombreux systèmes et philosophies radicalement différentes qui méritent d'être présentés ici pour mieux appréhender la mise en place du réseau.

3.1.1.1. *Le système d'information hospitalier (SIH)*

Selon [94], un SIH a pour objectif de structurer les informations et la gestion de celles-ci dans un environnement hospitalier, en accord avec les rôles des différents acteurs de la structure hospitalière. Le but d'un SIH est alors de permettre la gestion de ce qui est la fonction d'un hôpital, c'est-à-dire en tout premier lieu la gestion des patients, des soins ; mais aussi la gestion administrative et économique tout en respectant les dispositions légales.

Le défi d'un SIH est aussi de permettre d'amasser toute l'information produite en son sein et de la structurer de façon à permettre sa restitution en temps et lieu au personnel autorisé dans une forme la plus facilement exploitable.

3.1.1.2. *Le système d'information de santé (HIS)*

L'acronyme, HIS, pour *Health Information System*, désigne un système d'information qui fournit les bases d'un système permettant de recentrer la gestion des soins au niveau du patient. Les données médicales qui étaient centralisées au niveau de l'hôpital sont alors dispersées à travers un réseau distribué de centres de soins.

L'objectif de cette approche, outre la facilitation de la prise en charge médicale des patients est aussi de réduire les coûts d'intégration de cette vision de la gestion médicale centrée sur le patient.

3.1.1.3. *Le dossier médical informatisé*

Le dossier médical informatisé, ou *Electronic Record*, désigne une représentation sous forme électronique d'un dossier médical patient. Toujours d'après [94], à l'origine du dossier médical informatisé se trouvait une grande hétérogénéité, souvent à l'échelle hospitalière (SIH) de la conception de ces ensembles d'enregistrements.

Le contenu d'un dossier médical informatisé n'est cependant pas une chose triviale à définir. Ce dossier contiendra en priorité l'exhaustivité des données générées tout au long de son parcours de soin.

Le principe même d'un enregistrement informatisé est qu'il puisse facilement être intégré dans les bases de données hospitalières. L'idéal est que chaque enregistrement soit généré

électroniquement afin d'être directement intégrable dans le SIH, ce qui facilitera sa prise en charge administrative et comptable.

De façon plus pratique, le dossier médical peut être aussi considéré comme un environnement logiciel issu des technologies Internet qui permet aux personnes autorisées d'accéder en tout temps et tout lieu aux données concernant tout patient. D'ailleurs, lorsque le dossier médical est accessible à la fois en consultation et en modification à un ensemble d'acteurs médicaux on parle de dossier médical partagé ; ce qui représente un enjeu majeur de société à l'heure actuelle.

Selon l'ISO, un dossier médical partagé *Shared Electronic Health Record* [95] est défini ainsi :

"It will consist of a range of health organizations and clinicians attended by the patient/consumer on a regular or episodic basis. This will typically include one or more primary care clinicians, specialist clinics [...], hospitals, allied health professionals and alternative/complementary practitioners"

L'équivalent de l'ISO à l'échelle française, l'AFNOR¹, a mis en place une commission de normalisation pour l'informatique de la santé [96]. Cette commission s'inspire justement de ce document, entre autres, pour élaborer une norme nationale encadrant l'échange de données médicales et la création de dossier médicaux partagés. Ces documents ne seront disponibles qu'en 2012 pour les définitions générales et 2013 pour le modèle fonctionnel de dossier médical partagé, et cela malgré la mise en place fin 2010 du DMP en France avec un déploiement prévu tout au long de l'année 2011 [97].

3.1.2. Comparaison avec les systèmes existants

Les systèmes existants de partage et d'échange de données médicales en France sont souvent de la même nature, c'est-à-dire une architecture client-serveur étendue, suivant le modèle déclaratif présenté en [1.3.6.2]. La [Figure 15] représente schématiquement le fonctionnement bilatéral d'un modèle client serveur. Pour reprendre l'exemple de l'exploitation des données sur le cancer, des clients (médecins, hôpitaux, laboratoires) approvisionnent par un système déclaratif un système centralisé (INVS). Ces données sont traitées puis mises à disposition de clients (INVS, INSEE).

Avantages du système

Cette architecture possède néanmoins des avantages. Le fait de centraliser les données permet de faciliter leur prétraitement (nettoyage, standardisation, homogénéisation, présentation). De plus, le contrôle de l'intégrité des données est assuré par ce serveur central car les moyens de réplication et de sécurité ne sont à déployer qu'en un seul site.

Du côté client-exploitation, l'interrogation ne se fait qu'en un seul point. Ainsi, la collecte est largement facilitée par cet aspect. Le plus souvent, ces clients peuvent même disposer d'une copie locale des données nécessaires à leur traitement (sans information nominative) pour réaliser des études épidémiologiques.

¹ Agence Française de NORmalisation

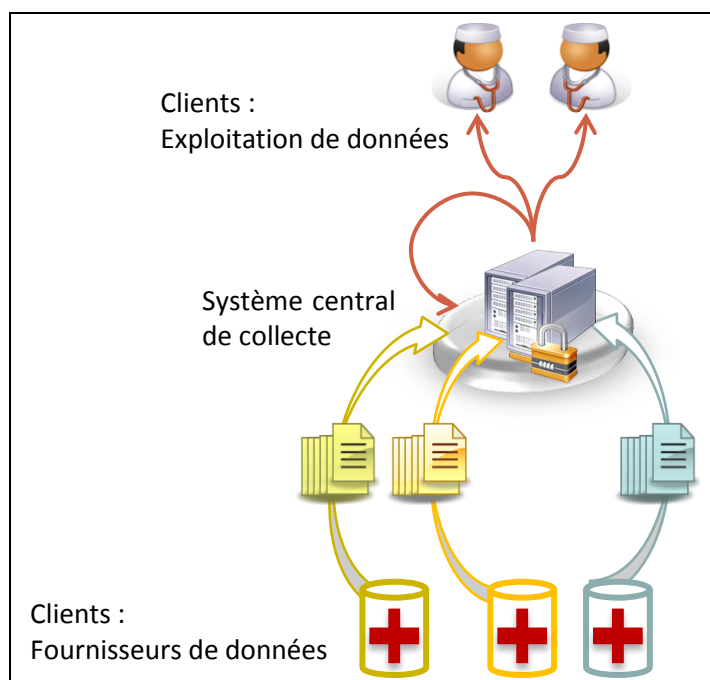


Figure 15 - Système de collecte

Inconvénients

Ce modèle client-serveur étendu a cependant ses défauts. Tout d'abord, les fournisseurs de données doivent au préalable préparer les données avant de les transmettre sur le serveur central (voir [1.3.6.2]), il leur en coûte un certain temps de collecte, d'analyse et de présentation des données (voir [Figure 16]) qui peut varier en fonction de la difficulté d'adaptation au système.

Le deuxième inconvénient est aussi le manque de contrôle sur les données : toute information envoyée au serveur central est définitivement perdue. L'exploitation ultérieure des données n'aura aucune idée de la source de celles-ci et les recours sont difficiles en cas d'erreur sur les données.

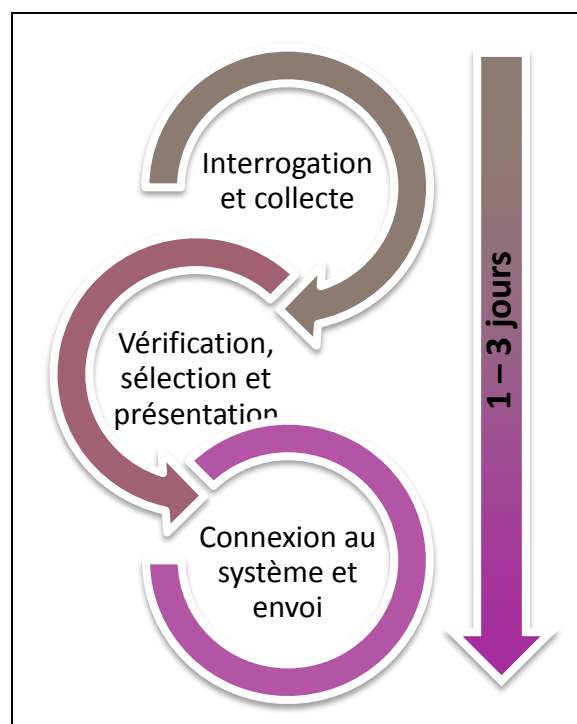


Figure 16 - Séquence d'envoi déclaratif de données

Ensuite, les délais entre la collecte et la présentation des données au niveau du service central peuvent prendre entre quelques mois et plusieurs années.

Ces délais évincent définitivement toute idée d'analyse ou de surveillance épidémiologique en temps-réel.

3.1.2.1. Les registres des cancers

Dans les départements où existent un registre (voir [1.2.2.2]), tous les praticiens ont pour obligation de déclarer tout cancer décelé dans la zone géographique concernée. Il suffit au médecin de se connecter au système hébergeant le registre et de renseigner ses observations. Les temps de réponse sont alors de l'ordre de 2 à 5 jours ouvrés.

L'inconvénient majeur d'un registre est son manque d'exhaustivité sur le territoire national. De plus, l'utilisation des données est uniquement réservée à l'épidémiologie. Par ailleurs, les comptes rendus de cancers bénins, en cas de dépistage, ne sont pas collectés (seuls les cancers avérés le sont).

3.1.2.2. Les données de mortalité

A des fins épidémiologiques il peut aussi être intéressant d'accéder à des données de mortalité, qui peuvent être collectées de deux façons différentes, soit par déclaration de décès ou par le biais des pompes funèbres. Les délais sont parmi les plus courts de toutes les sources de données médicales, de l'ordre de 2 à 7 jours ouvrés.

3.1.2.3. Les CRISAP

Les CRISAP établissent leur collecte sur un système déclaratif, par les pathologistes eux-mêmes. Ils doivent pour cela interroger leur système d'information, souvent grâce à un développement de la société éditrice de leur logiciel de gestion, formater ensuite cette requête au format CRISAP, puis l'envoyer au serveur centralisé qui se charge de son exploitation.

La disponibilité des données dépend de la bonne volonté du pathologiste d'exporter ses comptes rendus, bien qu'il soit possible d'automatiser la requête. La participation non-pérenne des pathologistes peut être préjudiciable à l'exhaustivité des données.

3.1.2.4. Le S MSC

Le Système MultiSource Cancer [98] est un projet porté par l'InVS pour recueillir des données cancer provenant des sources ALD, PMSI et ACP¹. La cohésion des sources de données est assurée par la création d'un nouvel identifiant commun au système central. La sécurité des communications est assurée par un mécanisme d'anonymisation nommé FOIN² proposée par la CNAM-TS³. Cette fonction d'anonymisation n'est cependant pas sans inconvénient car les procédures de correction d'erreur sont difficiles [99].

Qualifié de système lourd et coûteux, les développements n'ont toujours pas abouti et l'InVS semble avoir abandonné son implémentation, préférant se greffer sur les systèmes proposés par l'ASIP.

¹ Anatomie et Cytologie Pathologique

² Fonction d'Occultation d'Identifiants Nominatifs

³ Caisse Nationale d'Assurance Maladie des Travailleurs Salariés

3.1.2.5. Les recommandations de l'ASIP-Santé

L'Agence des Systèmes d'Information Partagés de Santé [5] a la charge de publier un ensemble de recommandations à destination des concepteurs de systèmes qui gèrent de l'information médicale. Un bon nombre d'entre elles ont déjà été proposées et d'autres sont en cours de validation.

Le rôle de cette organisation est principalement d'assurer l'interopérabilité des systèmes d'information médicaux [100], et cela en respectant la législation française. Elle a ainsi publié tout un cahier des charges sur la DMP-Compatibilité [101] et un ensemble de recommandations sur la sécurité, l'analyse des risques, la gestion des données d'identification et aussi l'hébergement de ces données.

Ainsi, sur le territoire français il est primordial de se conformer à ces recommandations pour avoir la certification nécessaire pour la communication avec les autres systèmes informatiques de santé.

Ces directives étant récentes, les réalisations en conformité avec les recommandations ne sont pas nombreuses malgré le lancement officiel du DMP en décembre 2010 [97]. De plus, pour le moment, seuls une douzaine d'hébergeurs de santé, principalement privés, ont reçu leur accréditation par l'ASIP pour des données médicales [102].

On peut notamment retenir, pour les besoins du projet, que l'utilisation des cartes de professionnel de santé est très fortement recommandée pour les phases d'authentification des utilisateurs, pour le chiffrement des communications ainsi que la signature électronique des actes médicaux.

3.1.2.6. Le Dossier Communicant de Cancérologie (DCC)

Dans le cadre de l'implémentation du DMP, le DCC [103] s'est créé afin de répondre à l'action 18.3 du plan cancer 2009-2013 [104] « Partager les données médicales entre les professionnels de santé » supervisé par l'INCa. L'objectif du DCC est alors de répondre à ce besoin d'échange de données dans le cadre du cancer.

Le DCC a pour mission principale de « faciliter la coordination des soins entre les professionnels de santé et la continuité des prises en charge des patients ». Il est clairement orienté à un usage clinique et devra servir à véhiculer l'information sur la maladie du patient durant toutes ses phases de traitement de sa pathologie.

Les travaux de mise en œuvre du DCC ont commencé en 2010 et prévoient un fonctionnement généralisé en 2013. Les étapes et écueils sont nombreux et ont été identifiés dans les phases expérimentales du projet. Les freins recueillis lors de cette phase sont de plusieurs natures :

- identification du patient ;
- utilisation de la CPS en milieu hospitalier ;
- problèmes de sécurité ;
- recueil du consentement du patient ;
- difficultés d'interfaçage avec les SIH ;
- manque d'interopérabilité.

A terme, le DCC devrait devenir une composante du DMP dédiée à l'oncologie.

3.1.3. Standards des dossiers médicaux partagés

Un grand nombre d'instituts de standardisation, au niveau national, européen ou international ont proposé différents standards de communication et de description des dossiers médicaux partagés, avec toujours le même objectif de faciliter l'interopérabilité, et ce, bilatéralement pour la collecte et la restitution de données.

3.1.3.1. openEHR

OpenEHR est une spécification d'un dossier médical partagé. Il est défini et promu par une association éponyme qui a pour objectif d'améliorer le système de soin en le rendant plus interopérable et durable. La spécification est fortement issue de standards de renom comme l'ISO ou HL7 [2].

Le principe de base de la spécification est de faire une scission (appelée archétype) entre la description au niveau système d'information, durable et solide et l'archétype spécifique à l'application, qui peut changer au fil du temps.

Selon openEHR [105], un archétype est une expression informatisée du contenu d'un domaine précis. Un archétype est en quelque sorte un modèle régi par des contraintes issues du modèle d'information (*openEHR reference model*). Ils sont créés de façon à être les plus réutilisables possibles, bien qu'ils puissent être spécialisés pour prendre en compte des particularités locales.

3.1.3.2. Health Level 7

HL7 définit un standard de communication pour l'échange clinique de messages. Il peut être utilisé pour l'administration comme pour la transmission de données, l'imagerie médicale ou la prescription. Un ensemble de types de données, messages et événements sont définis dans cette norme. La version 3 du standard s'appuie sur XML pour la description des messages, ce qui renforce un peu plus son caractère interopérable.

Le « *Reference Information Model* » (RIM) issu d'HL7 est la base de toute structure et de modèles de données développés pour HL7 V3. Le RIM est un modèle abstrait exprimé sous forme de diagrammes UML¹ avec des extensions spécifiques au HL7.

La [Figure 17] montre un exemple de spécification proposée dans RIM pour la gestion des rôles dans un système d'information médical, l'ensemble des spécifications est accessible en [106].

¹ Unified Modeling Language

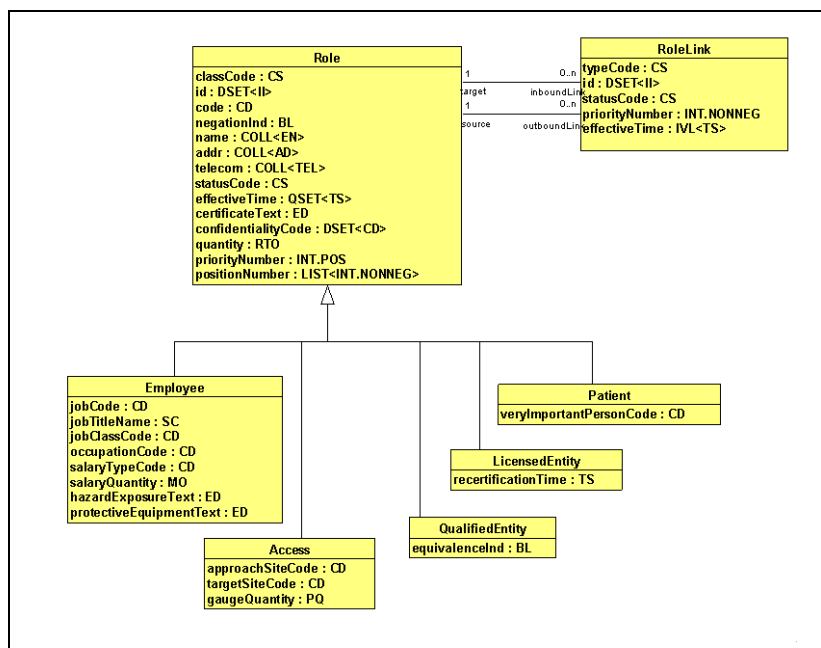


Figure 17 - Exemple de diagramme UML - Gestion des rôles par HL7-RIM - Source HL7.org

RIM sert aussi de modèle d'information au *Clinical Document Architecture* (CDA), développé aussi par l'organisation HL7. CDA se contente de définir le format des documents cliniques pour l'échange entre acteurs médicaux, ce n'est donc pas un modèle de dossier médical partagé mais il a vocation à en faire partie. Dérivé du HL7 v3, CDA utilise XML pour définir les données auxquelles peuvent s'ajouter un ensemble d'objets externes tels des images, vidéos, sons, etc. CDA constitue en cela la réponse stratégique d'HL7 pour l'interopérabilité du dossier médical partagé.

```

- <ClinicalDocument>
- <structuredBody>
- <section>
  <text>...</text>
  <observation>...</observation>
- <substanceAdministration>
  <supply>...</supply>
</substanceAdministration>
- <observation>
  <externalObservation>... </externalObservation>
</observation>
</section>
- <section>
  <section>...</section>
</section>
</structuredBody>
</ClinicalDocument>

```

Figure 18 - Document minimal HL7-CDA

La [Figure 18] présente une structure minimale d'un document XML rédigé suivant la spécification CDA.

3.1.3.3. Normes Européennes et ISO

Le Comité Européen de Normalisation [107], propose un standard de description des EHR. Ainsi, le *Technical Committee 251* [108] propose un standard pour l'informatique médicale, avec un volet important sur les dossiers médicaux partagés, avec toujours le même objectif central de faciliter l'interopérabilité des systèmes. Le standard est maintenant adopté par l'ISO sous la forme de deux

textes : ISO 18308 – « *Requirements for an Electronic Health Record Reference Architecture* » et ISO/DTR 20514 – « *Electronic Health Record Definition, Scope and Context* ».

Le premier propose une description des contraintes que présente une architecture de dossier médical partagé, fruit d'un long travail de collecte au travers des experts internationaux du domaine. La deuxième norme : ISO/DTR 20514 propose une classification des SEHR et en fournit un ensemble de définition des catégories et caractéristiques.

Le standard européen CEN13606 [109] quant à lui va beaucoup plus loin dans les règles d'implémentation des EHR. Ce document est scindé en 5 parties :

- *EHR Reference Model (02/2007)*

Fournit un modèle générique et compréhensif pour la partie communicante des EHR avec les systèmes existants.

- *Archetype Interchange Specification (07/2007)*

Approche orientée contrainte pour la définition de concepts d'objets cliniques. Issu d'HL7 et d'openEHR.

- *Reference Archetypes and Term Lists (02/2008)*

Modèle de conversion entre archétypes, compatible openEHR et HL7 RIM.

- *Security (03/2007)*

Modèle de partage de contrôle d'accès, recueil de consentement pour les communications.

- *Interface Specification (en cours)*

Spécification des messages et services pour la communication des EHR.

3.1.3.4. Relations entre ces normes

Schloeffel et al [110] ont proposé une décomposition en domaines [Figure 19] entre les différentes standardisations et systèmes de dossier médical partagé. Ainsi, openEHR s'appuie sur les standards CEN13606 et HL7CDA et peut interagir avec toute entité prenant en charge HL7 RIM.

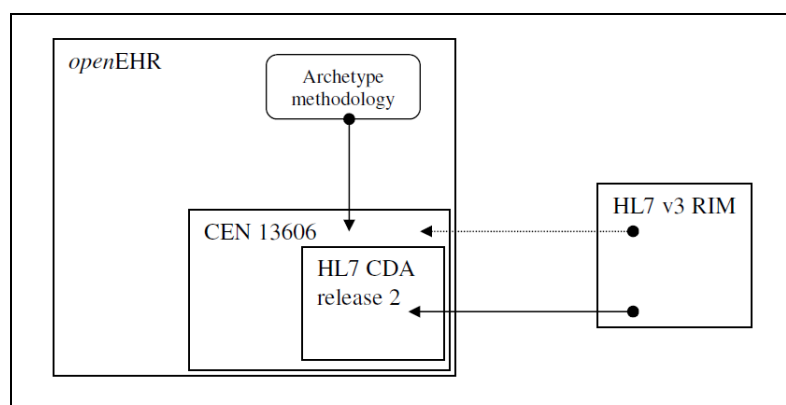


Figure 19 - Relations openEHR/CEN/HL7

3.2. LES TECHNOLOGIES INNOVANTES DE GRILLES POUR RSCA

Avant de décider de la configuration exacte de l'infrastructure de grille à mettre en place, il est important d'avoir une vision globale des mécanismes et des outils à mettre en œuvre pour l'échange sécurisé de données médicales afin de l'appliquer à RSCA. Une présentation générale de la grille EGEE, d'où est originaire l'intergiciel gLite est faite pour ensuite, en [3.2.3] l'appliquer pour RSCA.

3.2.1. La grille EGEE-EGI

Le projet EGEE [46] (Enabling Grids for E-Science) fédère une communauté de chercheurs issus de plus de 30 pays différents. L'objectif est de proposer une infrastructure commune de grille informatique pour les scientifiques. Succédant au projet DataGrid en 2004 [44], EGEE a duré 6 ans (2004-2010) sous forme de 3 projets successifs (EGEE I, II et III) financés par l'UE à hauteur de 30M€, voir [Figure 20]. Depuis 2010, EGI [47] a maintenant repris le flambeau, recadrant la gestion à une échelle nationale en assurant la maintenance, le développement et la promotion de l'outil.

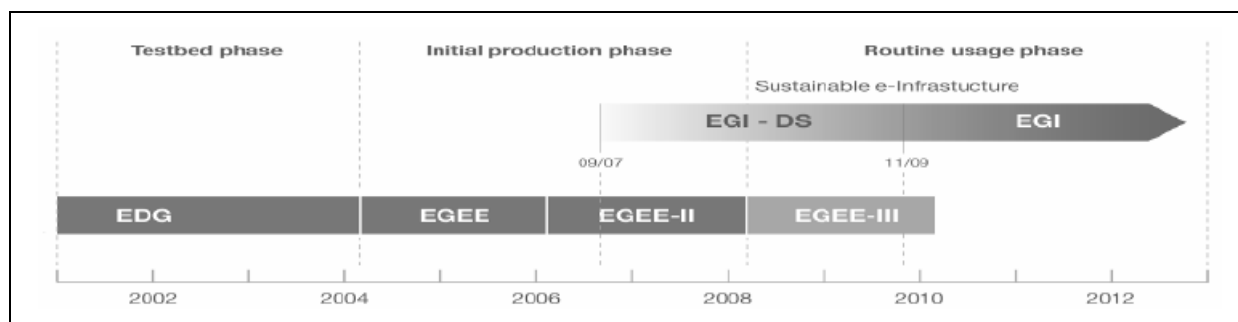


Figure 20 - Chronologie DataGrid - EGEE - EGI

La grille EGI est maintenant l'infrastructure de grille de production à l'usage de la communauté scientifique la plus large au monde. Selon Gstat [111], au début du projet EGI en 2010, on estime à 155000 le nombre de CPU logiques (disponibles 24/7) couplés à un espace de stockage de 64PB¹ dans la grille EGEE.

3.2.1.1. Fonctionnement, organisation

Le fonctionnement d'EGEE a quelque peu changé avec la transition vers EGI. Là où EGEE centralisait la gouvernance à un niveau global, EGI a mis en place une hiérarchie plus subtile. Techniquement, plusieurs NGIs² sont rattachées à EGI. Les NGIs ont pour mission de tenir en état de fonctionnement une grille nationale tout en se conformant à la politique globale EGI.

En France, la NGI a été créée sous la forme du Groupement d'Interêt Scientifique (GIS) France-Grilles [112]. Il est piloté par l'institut des grilles du CNRS dont l'actuel directeur est Vincent Breton. Il est formé d'un consortium public regroupant le CEA, la Conférence des Présidents d'Universités (CPU), le CNRS, l'INRA, l'INRIA, l'INSERM et l'infrastructure RENATER.

¹ 1PetaByte=10¹⁵ bytes

² National Grid Initiative

La [Figure 21] montre l'organisation adoptée entre EGI et les diverses grilles nationales dans les différents pays. EGI proposera une liste de tâches qui devront être réalisées par chaque NGI. Par ailleurs, chaque grille nationale sera administrée de façon locale, avec les tâches qui lui sont propres.

Cette infrastructure a l'avantage de laisser les pays membres maîtres de leur infrastructure et de leur maintenance. Le rôle d'EGI, bien que non-technique, est primordial pour assurer la cohésion de l'ensemble.

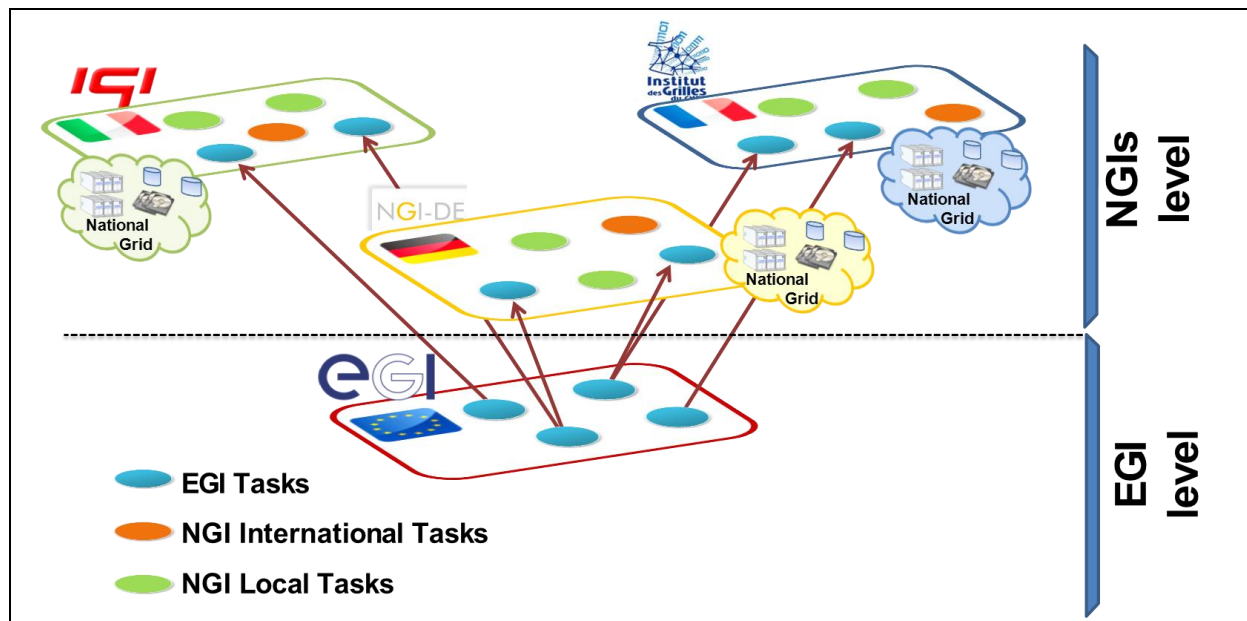


Figure 21 - Organisation des tâches EGI-NGI

EGI a apporté à la grille EGEE un modèle organisationnel géré au niveau national, ce qui, jusque là pouvait lui faire défaut sur la gouvernance. La pérennité de l'infrastructure est davantage garantie avec ce modèle.

3.2.1.2. Développement et intergiciel

La grille EGI s'appuie naturellement sur l'intergiciel gLite, qui a été en grande partie développé lors des successifs projets EGEE. L'ensemble des technologies présentées en [3.2.1] sont donc représentées largement au sein des différents services et sites d'EGEE.

Le développement de la grille EGI, grâce à son modèle ouvert, permet à tout organisme de recherche de rejoindre, collaborer et contribuer aux différents projets de la grille. Ainsi, chaque utilisateur des pays membres peut s'adresser à une structure tutelle de son pays pour accéder aux ressources de la grille.

3.2.1.3. Grilles dérivées : grilles privées, dédiées, réutilisation des technologies

Toujours grâce au modèle de développement open-source de l'intergiciel gLite, les réutilisations des couches logicielles issues du développement d'EGEE sont nombreuses.

Les technologies de grilles ont de nombreux avantages pour résoudre les problèmes d'accès aux données ou de calcul distribué. La prise de conscience de ces possibilités a soulevé un intérêt grandissant pour ces technologies. Cependant les développements nécessaires à la réalisation d'une telle infrastructure pouvaient rebuter les personnes intéressées.

Ce n'est qu'avec l'avènement de la grille EGEE et de son intergiciel gLite, diffusé sous licence libre, que les développements de grilles privées ont débuté. En effet, les technologies issues d'EGEE, qui fournissent un ensemble de services d'information, de monitoring, de gestion des tâches, le tout librement réutilisables, a permis de les démocratiser.

C'est ainsi que des grilles privées, souvent couplées à une problématique précise sont nées de la réutilisation des technologies gLite. Le plus souvent ces grilles dédiées sont liées au domaine biomédical, qui a plus particulièrement besoin d'un espace plus confidentiel. Le projet pionnier dans ce domaine été MammoGrid [73]. Ont suivi Health-e-child [89], neuGRID [113] ou encore e-nmr [114].

3.2.1.4. Limitations

L'intergiciel gLite propose une infrastructure logicielle adaptée pour les scientifiques qui veulent accéder à un grand nombre de ressources matérielles sans se soucier des problèmes liés à la distribution géographique des sites. Cependant, certaines applications ne sont pas du tout adaptées à ces technologies car elles présentent trop de couplage entre les tâches. Une succession de tâches courtes mais nombreuses n'aura pas l'effet escompté sur la réduction du temps global d'exécution. C'est en partie expliqué par les délais des files d'attente sur les nœuds de calcul qui sont souvent supérieurs au temps d'exécution pour de petites tâches. Un juste équilibre est à trouver entre nombre de tâches parallèles et durée de ces tâches.

Le projet HOPE [85] a justement étudié l'impact du nombre de jobs par rapport au temps total de simulation d'une tâche de simulation Monte Carlo utilisant le logiciel GATE a été montré [115, 116]. On voit, sur la [Figure 22] que le temps total d'exécution diminue fortement entre 1 et 20 jobs mais que cette tendance s'inverse lorsque l'on passe au-delà des 20 jobs.

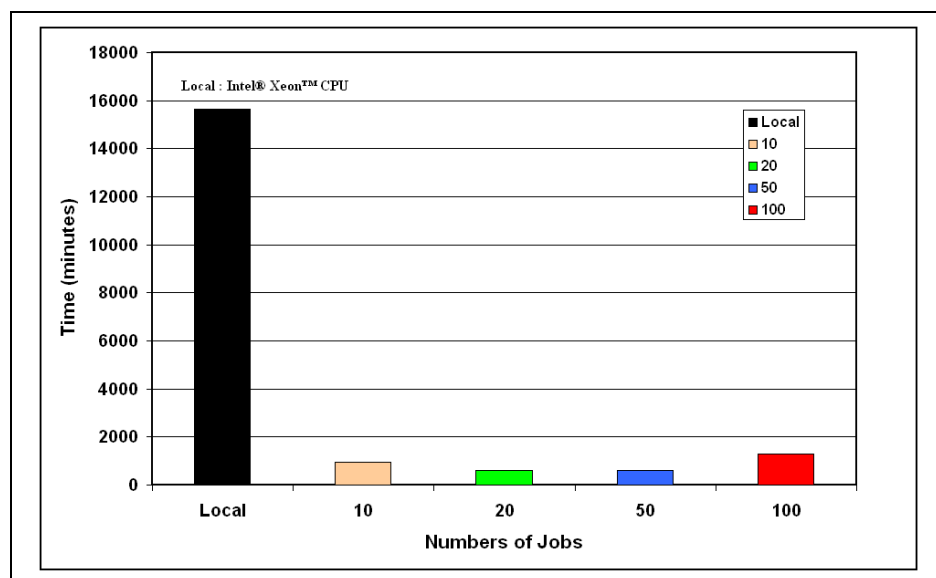


Figure 22 - Evolution du temps de calcul d'une simulation par rapport au nombre de jobs sur grille - source [116]

Une modélisation plus détaillée de ce phénomène de latence des grilles a été présentée en [117] sur la grille EGEE, notamment grâce à une modélisation poussée des lois qui régissaient ces délais, avec une validation sur des mesures expérimentales.

Il s'agit alors, d'adapter le plus précisément possible la durée de ses jobs par rapport à leur nombre si le processus calculé permet cette flexibilité.

Les limitations de gLite sont clairement indiquées ici, car tous les programmes ne sont pas « gridifiables » avec un gain¹ proportionnel au nombre de nœuds.

Dans un autre registre, l'intergiciel gLite reste complexe à utiliser pour une personne sans compétences en informatique. Il y a donc besoin de services de haut-niveau s'appuyant sur la grille pour fournir aux utilisateurs une réponse claire à leurs besoins.

3.2.2. L'intergiciel gLite adapté à RSCA

L'intergiciel, ou *middleware* gLite [118] est né pour le développement du projet EGEE [46]. gLite est fondé sur l'infrastructure open-source Globus [119] et son *toolkit* dont il reprend en partie les composantes. C'est une adaptation des outils d'informatique répartie issus de Globus à l'infrastructure conçue pour le démarrage du projet EGEE.

gLite est constitué de différents services qui sont pour chacun d'entre eux implémentés sous forme de logiciels.

Parmi les composants logiciels on peut distinguer plusieurs familles de produits : la gestion de données, le système d'information, de log et de supervision, la gestion des tâches (jobs), les services de sécurité, et les composants utilisateur, qui permettent l'interaction homme-grille. L'ensemble de ces composants sont présentés, téléchargeables et documentés sur le site de gLite [118].

3.2.2.1. Services gLite

gLite est organisé en différentes composantes, comme montré en [Figure 23]. Parmi les composants essentiels on distingue :

- le service d'autorisation et d'authentification qui assure une grande partie de la sécurité des utilisateurs ;
- les ressources, sous forme de puissance de calcul (clusters, superordinateurs, ...) ou d'espaces de stockage (ou base de données) ;
- le service « Logging and monitoring », qui se charge de suivre l'évolution de l'état des ressources, des jobs et des utilisateurs. Ce composant permet en outre de connaître l'historique de l'utilisation des ressources ;
- le service d'information, qui suit l'évolution des sites, de leur état (disponibilité) ;
- l'élément de gestion des données, qui synchronise et permet d'utiliser au mieux les ressources de stockage par les jobs des utilisateurs (entrées/sorties) ;
- le composant central de gestion de la charge de travail, qui orchestre les différentes demandes des utilisateurs et les adapte en fonction d'un ensemble d'informations mises à disposition par les autres services.

¹ Ou speed-up : mesure d'accélération d'un programme distribué comparé à sa version séquentielle

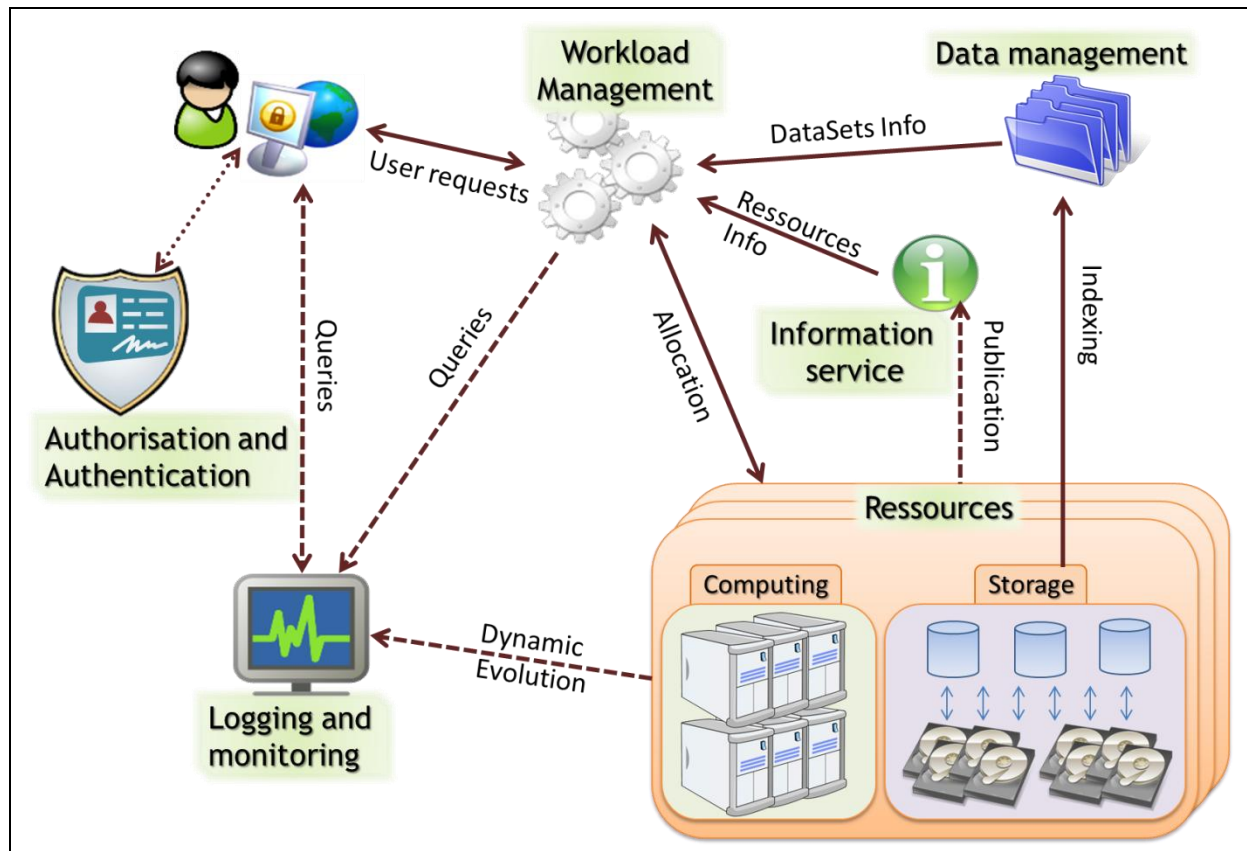


Figure 23 - Composants gLite

Cette décomposition en services permet aux grilles propulsées par gLite de les rendre les plus flexibles possible. De plus, certains services sont conçus pour être facilement dédoublables : en effet, une grille ne peut pas fonctionner en utilisant un unique système de gestion de la charge par exemple, ce qui en ferait un SPOF¹ bien trop critique.

Cependant certains composants ne peuvent que trop difficilement être répliqués. Afin de garantir les propriétés essentielles d'atomicité et d'intégrité de bases de données nécessaires à la sécurité au sein d'une même VO [1.3.4.4], les composants d'authentification et d'autorisation sont souvent concentrés en un seul site, tout comme la gestion de données. Seule une réplication intra-site est envisagée (alias DNS, machines et réseaux redondants).

3.2.2.2. Les services de sécurité : Public Key Infrastructure (PKI) et X.509

Au cœur de la sécurité de l'intergiciel gLite se trouve le système PKI [120], défini par la RFC3280 [121], largement repris par le *Grid Security Infrastructure* (GSI) ou *Globus Security Infrastructure* [122], qui offre des mécanismes cryptographiques de la forme clé privée/clé publique. Ces deux clés, mathématiquement indissociables, permettent en pratique que tout message chiffré avec l'une des deux clés ne puisse être déchiffré que grâce à l'autre.

Au niveau de la grille, ces mécanismes asymétriques permettent un haut niveau de sécurité des utilisateurs, des communications et aussi, surtout, de la certification des différents composants informatiques déployés tout au long de la grille.

¹ Single Point Of Failure

X.509 est la norme cryptographique régissant l'infrastructure PKI adoptée par gLite. Un certificat X.509, contrairement au logiciel PGP¹ [123] où n'importe qui peut signer le certificat d'un autre, X.509 repose sur un système hiérarchique d'autorités de certification qui peuvent valider et vérifier le certificat.

Une autorité de certification a pour mission de délivrer des certificats aux demandeurs. Ainsi, les utilisateurs doivent s'adresser à l'autorité de certification dont ils dépendent, sachant qu'il existe, au sein des différentes gouvernances une autorité primaire, dont des organismes (établissements de recherche, société spécialisées, banques, etc.) dépendent directement. Ainsi, pour un chercheur français, le CNRS a autorité reconnue pour délivrer des certificats.

Par ailleurs, un certificat n'a de valeur que si l'autorité l'ayant délivrée est reconnue par la partie exploitant ce certificat. Pour cela, il suffit d'ajouter le certificat racine de l'autorité de certification en question à la liste des autorités de certification ayant droit sur le service considéré. Cette mesure est primordiale car n'importe qui peut émettre un certificat.



Figure 24 - Exemple de certificat – Vue Kleopatra

Afin de renforcer la sécurité, le certificat est toujours chiffré selon le support où il est stocké. Il est alors nécessaire d'indiquer le mot de passe préalablement à l'utilisation du certificat.

La [Figure 24] montre un exemple d'un certificat, on distingue clairement le sujet, qui est la partie la plus importante et l'émetteur du certificat, qui garantit sa conformité. Les dates de validité sont aussi mentionnées, le plus souvent pour une durée d'un an.

Le principe de certificat est exactement le même pour la sécurisation des sites sur internet avec le protocole *https*. Le site émettant la page web sécurisée propose son certificat pour chiffrer la transmission des données. Le navigateur du client va pouvoir vérifier la validité de ce certificat en comparant sa signature avec celles dont il reconnaît l'autorité. Ces autorités racine sont moins d'une centaine, des sociétés telles que *Verisign*, *Thawte* ou encore *Kynectis* en France ont cette accréditation. La [Figure 25] présente sous forme d'un diagramme de séquence les différentes étapes permettant à un utilisateur d'établir une connexion certifiée et sécurisée avec un serveur en utilisant les certificats et autorités de certification.

¹ Pretty Good Privacy

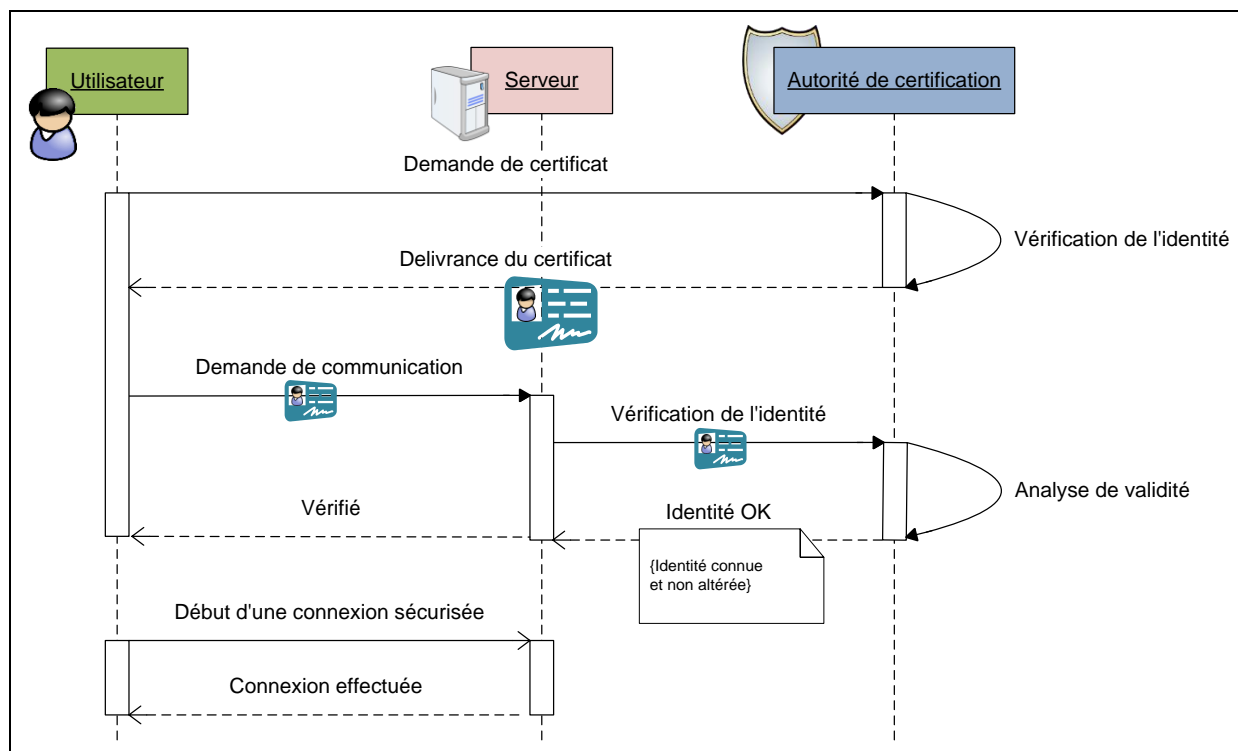


Figure 25 - Diagramme de séquence : Etablissement d'une connexion par certificat

Les Certificat Revocation Lists (CRL)

Les CRLs sont un élément important de l'infrastructure PKI. Publiées de façon périodique, ces listes constituent un ensemble de certificats émis par l'AC en question, qui ne sont plus valides. Outre l'expiration naturelle, certains certificats doivent être invalidés avant cette date limite. Cette révocation est nécessaire dans plusieurs cas de figure, comme la clé compromise, la suppression de certains privilèges. Cette liste est résumée dans la RFC3280 [121], voir [Figure 26].

```

ReasonFlags ::= BIT STRING {
    unused                (0),
    keyCompromise         (1),
    cACompromise          (2),
    affiliationChanged     (3),
    superseded             (4),
    cessationOfOperation  (5),
    certificateHold        (6),
    privilegeWithdrawn     (7),
    aACompromise          (8) }
  
```

Figure 26 - RFC3280 : Motifs de révocation des certificat

Compte tenu du caractère urgent de certains cas, il est nécessaire que le serveur d'authentification procède à une mise à jour de façon régulière et automatique de ces CRLs.

3.2.2.3. Sécurité : VOMS

VOMS est un système qui gère l'ensemble des autorisations des composants de grille. Il permet surtout aux utilisateurs de s'authentifier sur la VO et bien entendu faire que cette authentification ait validité sur l'ensemble de l'infrastructure voir [Figure 27].

La collaboration entre VOMS et les Autorités de Certification (AC) est très étroite et constitue la clé de voûte de toute la sécurité du système. Toute demande d'authentification via certificat passe par une vérification

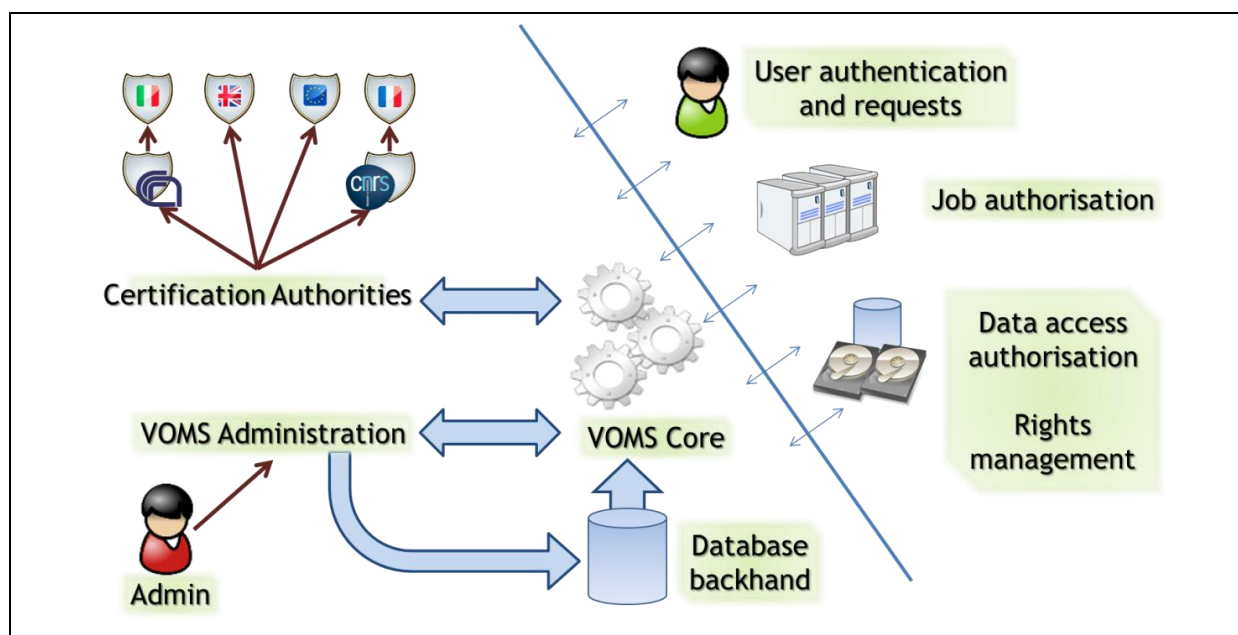


Figure 27 - Architecture de VOMS

Dans VOMS, le système central stocke toutes les informations nécessaires au stockage des utilisateurs, composants de grille, organisations virtuelles et les différents droits de chacun dans une base de données. Une interface web, voir [Figure 28] (ou client en service web) permet d'administrer VOMS, paramétrer les autorisations, ajouter/supprimer des accréditations et consulter les journaux d'accès, d'erreur, etc. Depuis cette interface on distingue clairement l'utilisateur en question ①, l'autorité ayant délivré le certificat ② ainsi que les différents droits dont il dispose ③.

Figure 28 - Interface d'administration de VOMS

Par la suite, l'ensemble des éléments d'une grille s'adressent à VOMS pour vérifier la validité d'une demande d'un utilisateur. Ainsi, la gestion des tâches et des données est parfaitement sécurisée.

VOMS est aussi l'élément qui, lors de l'authentification d'un utilisateur sur une VO, va lui créer un proxy grille, issue de son certificat, qui lui permettra d'utiliser les ressources de la VO pour une durée comprise entre quelques heures et quelques jours (24h le plus souvent). Le proxy est alors valide sur l'ensemble de la VO pour cette durée, que ce soit pour la gestion de données ou l'exécution de tâches. Cette durée limitée est un élément supplémentaire de sécurité, une personne parée de mauvaises intentions qui arriverait à intercepter le proxy, ou arriverait à s'authentifier sur une machine où un utilisateur a un proxy ouvert aurait un champ d'action limité dans le temps.

VOMS: Les groupes et les rôles pour la VO Sentinelle

Au sein d'une VO, les utilisateurs bénéficient de droits spécifiques. Il est ainsi possible de partager les droits d'accès aux fichiers des utilisateurs d'une même VO. Il existe aussi la notion de groupes, qui permet d'obtenir des droits spécifiques la VO (administration ou simple utilisateur par exemple). Des rôles peuvent aussi être attribués à des utilisateurs afin d'ajuster encore plus finement les droits.

Les groupes permettent aussi, en cas d'extension du réseau sentinelle, de définir des droits propres à une application sans devoir créer une nouvelle VO. Un groupe « cancer-depistage » au sein de la VO Sentinelle rassemble tous les acteurs ayant besoin de récupérer les comptes rendus pathologiques nominatifs et un groupe « cancer-epidemio » considèrera seulement les utilisateurs ayant besoin de lancer des requêtes statistiques.

En cas d'extension du réseau, comme mentionné dans les objectifs généraux du cahier des charges [2.1.3], un autre groupe peut voir le jour sur une autre application que le cancer. Pour cela, ce groupe sera créé dans la VO Sentinelle pour gérer ce nouveau champ d'application, voir [Figure 29].

Le système de gestion de données utilisé, AMGA, présenté ultérieurement en [3.2.2.8] peut alors récupérer l'information sur l'appartenance d'un utilisateur aux groupes constitués dans VOMS et contrôler l'autorisation à accéder aux données en fonction.

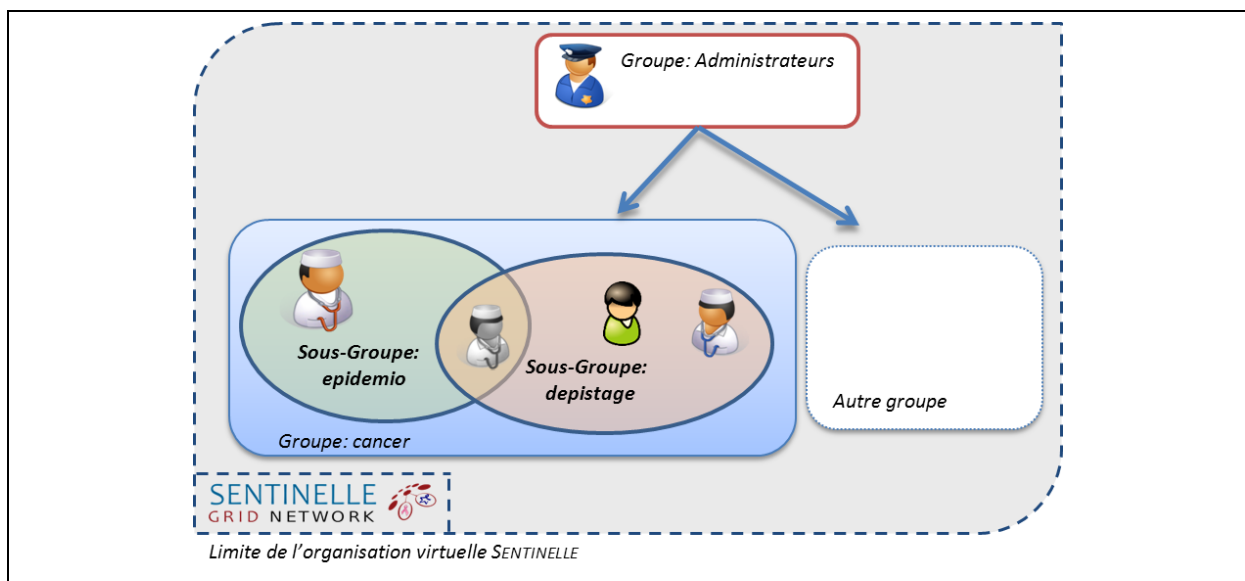


Figure 29 - Composition des groupes dans la VO Sentinelle

3.2.2.4. *L'interface utilisateur (UI)*

L'UI est le point d'entrée de tout utilisateur sur la grille. C'est à cette entité, souvent une machine connectée à un réseau haut débit, que les utilisateurs se connectent pour interagir avec la grille. La configuration d'une UI est largement documentée pour faciliter son déploiement [124].

Les systèmes d'exploitation officiellement supportés par gLite sont Scientific Linux 4 et 5 qui sont des dérivés de RedHat Enterprise Linux 4 et 5.

L'UI va permettre, entre autres, à un utilisateur de :

- s'authentifier sur la VO qu'il souhaite utiliser en créant un proxy ;
- connaître le statut et les différents sites accessibles, lister l'ensemble des sites capables d'exécuter son job en fonction du logiciel à exécuter ;
- gérer ses données : envoyer et récupérer des données sur la grille, les répliquer et les partager (gestion des droits) ;
- soumettre et superviser des jobs, récupérer les informations d'exécution.

3.2.2.5. *Le service d'information*

Comme son nom l'indique, le service d'information a pour objectif de recueillir les données nécessaires à connaître le statut de la grille. MDS¹ et R-GMA² sont les deux éléments de surveillance de l'état de la grille. MDS recueille les statuts des différentes ressources et s'occupe de publier ces statuts via un serveur appelé *resource-level BDII*³ [125]. Ces statuts sont agrégés via des serveurs relais au niveau site (*site level BDII*), puis au niveau global (*top-level BDII*) pour être ensuite dispersés aux clients. Techniquement BDII est une base de donnée de type LDAP⁴ qui implémente le schéma GLUE [126], qui fonde toute une description des ressources de grille sur une ontologie.

R-GMA se charge quant à lui du volet comptabilité, surveillance et publication d'informations aux utilisateurs. Il agrège et collecte les informations sur les jobs qui se sont exécutés sur la grille et permet d'en tirer des statistiques sur l'utilisation de celle-ci.

Les UI sont dotées de commandes spécifiques à l'interrogation des BDII, permettant ainsi à un utilisateur de connaître le statut des ressources, l'espace restant sur un SE ou l'état de la file d'attente sur un CE. Les sites peuvent aussi disparaître du BDII lors de maintenances programmées par exemple afin de prévenir les utilisateurs d'essayer d'utiliser ces ressources, ce qui pourrait être fatal pour un job.

3.2.2.6. *La gestion des tâches*

Les jobs, composants centraux du calcul sur grille, doivent être gérés par des dispositifs performants. L'objectif premier d'un job est qu'il s'exécute, et cela le plus rapidement possible compte tenu de l'environnement de la grille. Le composant de gLite délégué à la gestion des tâches est WMS, *Workload Management System* [127], qui se charge de répertorier les requêtes puis de les traiter de la façon la plus efficace possible, c'est-à-dire choisir le CE le plus approprié pour effectuer

¹ Globus Monitoring and Discovery Service

² Relational Grid Monitoring Architecture

³ Berkeley Database Information Index

⁴ Lightweight Directory Access Protocol

la tâche. WMS est en quelque sorte un méta-ordonnanceur de tâches au sein d'un environnement distribué.

Un job s'écrit à partir d'un fichier au format JDL, *Job Description Language* [128] dont un exemple est fourni en [Figure 30]. Ce fichier a pour but de décrire de manière précise l'objet du job, les données nécessaires en entrée et sortie. Il permet aussi de préciser un ensemble de pré requis sur la capacité mémoire, le nombre de cpu ou encore le logiciel qui doit être installé sur le nœud en question.

```
[
  Executable = "/bin/sh" ;
  Arguments = "./example.sh" ;
  StdOutput = "std.out" ;
  StdError = "std.err" ;
  OutputSandbox = {"std.out","std.err","result.dat"};
  RetryCount = 3 ;
  Type = "Job" ;
  JobType = "normal" ;
  InputSandbox = {"data/input"};
  requirements =
    ((other.GlueCEPolicyMaxCPUTime>102)) ;
]
```

Figure 30 - Exemple de fichier JDL

WMS a donc pour rôle de soumettre et de surveiller les jobs. Il laisse la possibilité à l'utilisateur de savoir à tout instant le statut de ses tâches au travers du BDII. La [Figure 31] récapitule, sous forme d'un diagramme de séquence, l'enchaînement possible des états d'un job sur la grille depuis la soumission à la récupération des résultats.

La communication entre WMS et BDII est très importante car c'est le statut de la grille au travers du système d'information qui permet à WMS d'orienter les demandes de job le plus précisément possible, en fonction des caractéristiques du job, ou de l'état des files d'attentes des différents CE.

Le rôle central de BDII est clairement exposé ici, car de mauvaises informations publiées dans le système d'information fausseraient le fonctionnement global de la grille. Par exemple si un CE ne met pas à jour l'état de sa file d'attente, il pourrait se voir surchargé de demandes de job ou au contraire boycotté par les demandes utilisateurs.

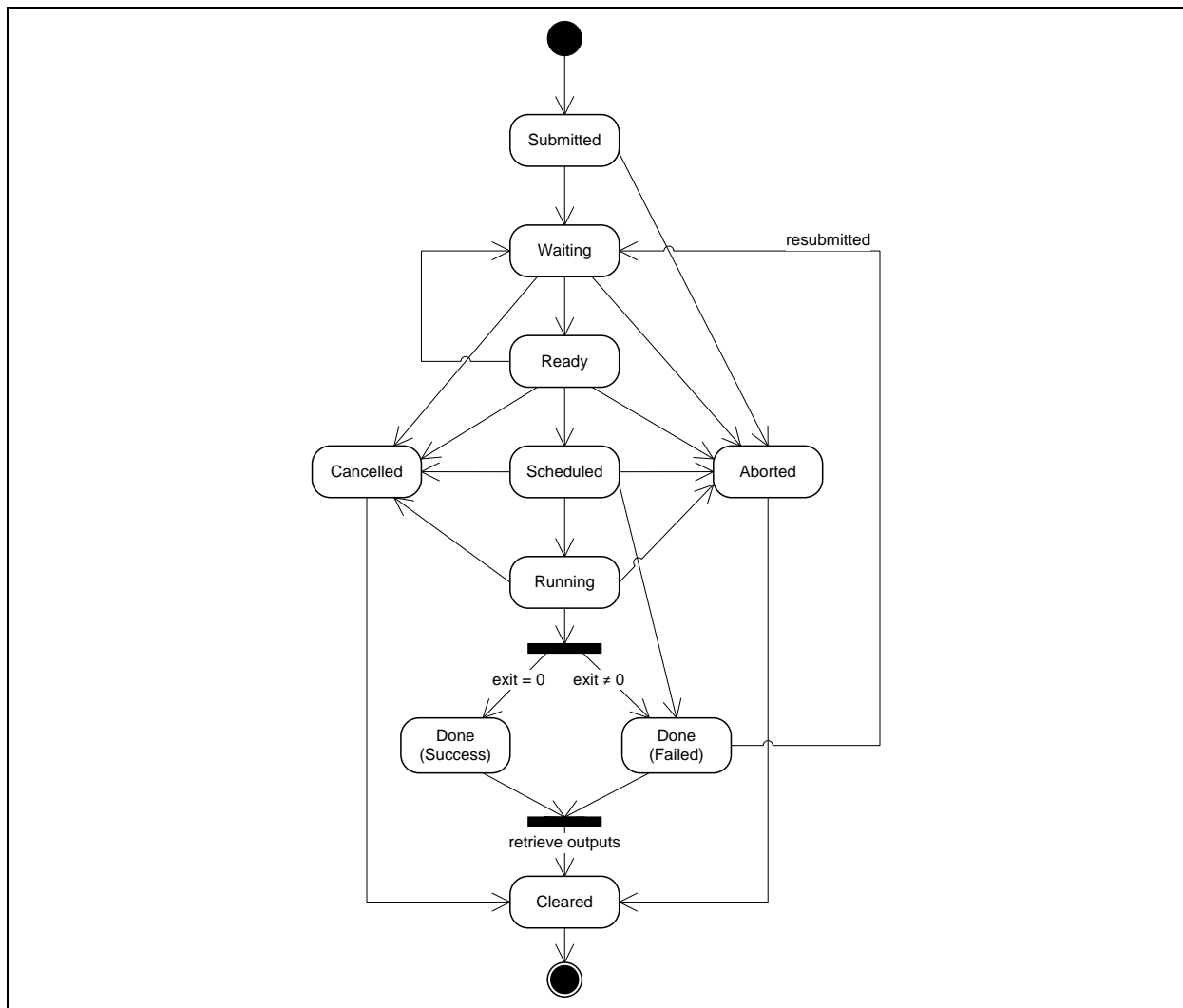


Figure 31 - Cycle de vie d'un Job

3.2.2.7. La gestion de données

gLite a été conçu pour que les utilisateurs puissent aisément envoyer des fichiers sur la grille, effectuer les opérations nécessaires sur ces données (analyse, traitement), créer de nouveaux fichiers pour stocker les résultats et les récupérer, tout cela le plus facilement possible. Le composant le plus utilisé pour effectuer ces opérations est gridFTP [129], un client FTP optimisé pour la grille, notamment en terme de débits afin de maximiser l'usage de bande passante mais aussi pour se conformer à la politique de sécurité en utilisant la couche de sécurité issue de Globus.

Par-dessus cette couche basse de transfert de fichiers a été développée une application à un niveau plus logique, baptisée Lcg File Catalog (LFC) [130], qui est en quelque sorte un méta-gestionnaire de fichiers adapté aux environnements de grille.

Le LFC met en œuvre trois entités principalement :

- Le GUID, pour *Globally Unique IDentifier*, qui identifie de façon unique un fichier
- Le LFN, pour *Logical File Name*, est un alias (nom) donné au fichier sur la grille. Il a pour format : `lfn:/grid/...`
- Le SURL, pour *Storage URL*, qui est l'adresse physique d'un fichier sur le SE

Le fonctionnement de LFC pour le stockage d'un fichier, est assez simple [Figure 32]:

- étape1 : L'utilisateur veut envoyer un fichier sur un SE de la grille. Il fournit un nom (LFN) de manière à l'identifier correctement ainsi qu'un SE de destination (SE2.in2p3.fr) ;
- étape2 : Le LFC reçoit cette demande et crée un identifiant unique (GUID) auquel il associe le LFN ;
- étape3 : L'utilisateur envoie physiquement le fichier au SE en précisant le GUID (SE2.cea.fr) et fournit un SURL (adresse physique du fichier sur le SE) ;
- étape4 : Le LFC mémorise l'association entre le GUID et le SURL.

Ces quatre étapes présentent le fonctionnement du LFC pour stocker un fichier. Par la suite des étapes supplémentaires peuvent être réalisées par l'utilisateur :

- la réplication : l'utilisateur souhaite répliquer son fichier sur le SE (SE1.in2p3.fr). Le LFC transmet cet ordre et le fichier est envoyé du SE2.cea.fr au SE1.in2p3.fr. Un nouveau SURL est alors enregistré pour le guid de ce fichier, qui est maintenant répliqué en deux endroits ;
- l'association d'un autre nom : L'utilisateur peut associer un autre LFN au GUID, une simple association est créée.

Par la suite, l'utilisateur, que ce soit à travers une UI ou via un job qui s'exécute sur la grille pourra, via le LFC, récupérer ces fichiers ou en stocker de nouveaux.

Les performances d'un tel système, qui est un point essentiel de la gestion de données sur grille, ont été discutées et étudiées [131], par rapport à une autre application similaire existant précédemment dans gLite : FiReMan¹ [132].

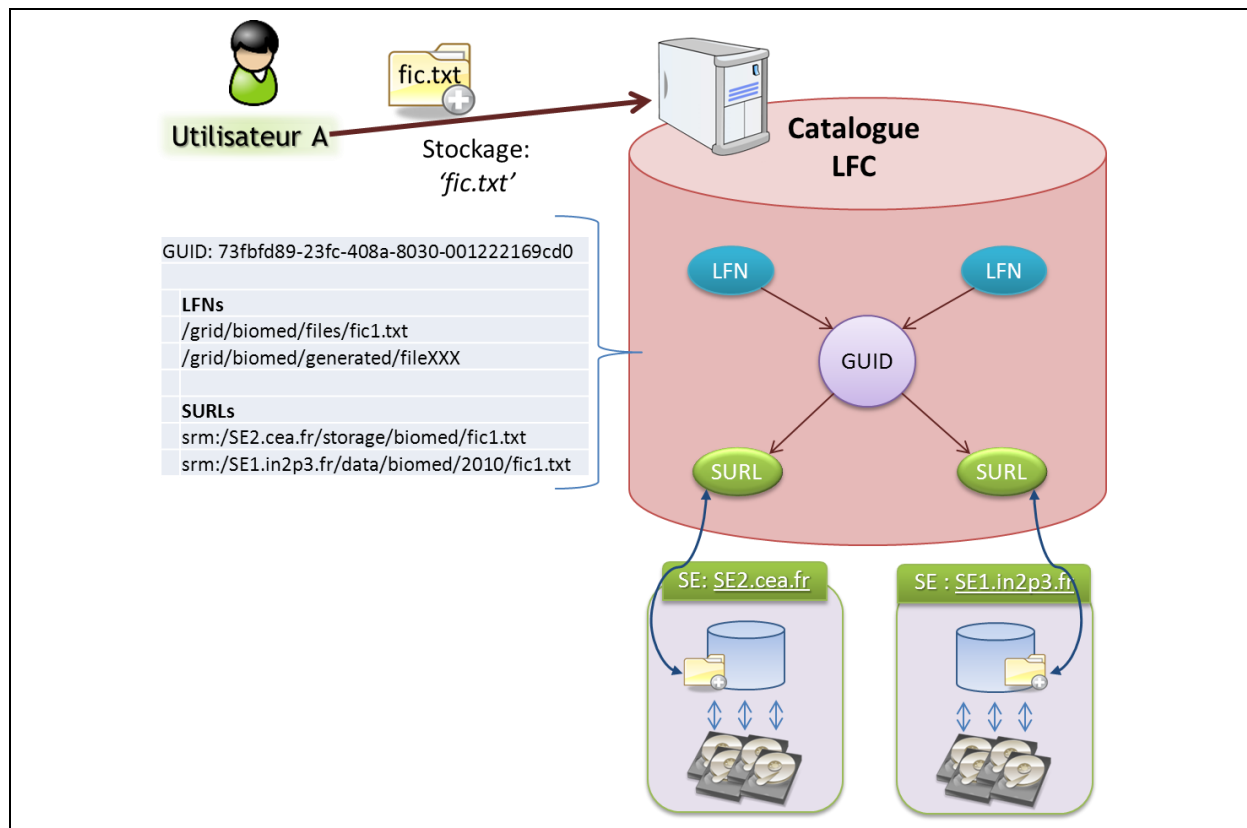


Figure 32 - Fonctionnement du LFC pour le stockage d'un fichier

¹ File Replication MANager

La gestion des données dans RSCA

La VO Sentinelle, dans l'application définie dans le cahier des charges en [2.4.2.3] sur l'échange des comptes rendus d'anatomie pathologique et en [2.4.2.4] pour l'accès aux données statistiques n'est pas très volumineuse sur l'utilisation des données. L'intégralité des informations peuvent être contenues dans une base de données. En effet, si on tient compte des statistiques des associations de dépistage [2.2.2.1], bien qu'elles doivent gérer sur la région environ 100000 personnes, moins de 1000 comptes rendus médicaux sont réellement à récupérer par an.

Cependant, dans un second temps, comme évoqué en [2.4.2.5], l'introduction de l'imagerie médicale pour le transfert des mammographies rendra l'utilisation de toute la pile de gestion des fichiers des grilles indispensable. Le volume des radiographies numériques ne permet pas de les stocker durablement dans une base de données.

3.2.2.8. AMGA

Généralités

AMGA, *ARDA Metadata Catalog Project* [133] est un catalogue de métadonnées pour les environnements de grille, et plus précisément adapté pour gLite qui maintenant l'intègre complètement.

Le besoin de ce type de logiciel pour la grille s'est vite rendu primordial car les systèmes de gestion des données comme LFC, présenté en [3.2.2.7] ne permettaient pas d'annoter suffisamment les fichiers stockés sur la grille. Le LFC se charge uniquement d'associer à un nom logique, un (ou plusieurs) emplacement(s) physique(s) sur la grille. Le plus souvent, la mise à disposition de fichiers sur la grille ne permettait pas à un autre utilisateur de les exploiter correctement, souvent par manque d'information sur le contenu du fichier.

Un catalogue de métadonnées permet d'associer à un élément un certain nombre d'informations supplémentaires *caractérisant* cet élément.

AMGA diffère aussi d'une base de données classique en proposant une interface arborescente aux données. La base dispose ainsi d'une racine et d'un ensemble de répertoires qui contiennent un ensemble d'entrées et d'attributs caractérisant ces entrées.

AMGA : techniquement

AMGA fonctionne comme une couche de sécurité au dessus d'une base de données relationnelle conventionnelle (type Oracle, MySQL ou PostgreSQL) via un connecteur ODBC¹. AMGA peut ainsi être utilisé de façon parfaitement identique à une base de données, seule la syntaxe diffère. A noter que depuis la version 2.0.0 d'AMGA, le support de SQL est complet, l'apprentissage de la syntaxe spécifique d'AMGA n'est dorénavant plus nécessaire.

D'un côté implémentation, des APIs² sont disponibles pour C++, JAVA et Python permettant ainsi, du côté client, de rendre accessible AMGA facilement.

¹ Open DataBase Connectivity

² Application Programming Interface

Les évolutions d'AMGA ont continué à rendre le logiciel de plus en plus attractif vis à vis de l'accès distribué aux données. En effet, AMGA intègre en plus d'un serveur de bases de données un système intégré de réplication entre sites. Ainsi, des instances distribuées du serveur sont capables, en un temps quasi-réel de synchroniser leur contenu, ou une partie de leur contenu. Cette fonctionnalité de réplication a un impact primordial sur le réseau sentinelle car il permet qu'un ensemble d'instances participent conjointement à une partie de l'architecture des données. De cette manière les requêtes sont naturellement distribuées sur tous les sites participants de manière transparente.

Toujours dans le souci de s'adapter le plus possible aux infrastructures réparties, avec potentiellement un grand nombre de connexions simultanées, AMGA s'est doté d'une interface très légère entre la base sous-jacente et son moteur, afin de s'approcher au plus des performances natives du SGBD. De plus, le serveur exposé est dorénavant multithreadé, ce qui le rend d'autant plus performant sur les machines actuelles en ayant un potentiel bien plus élevé en termes de nombre de connexions simultanées.

Parmi l'analyse initiale des besoins lors de la conception d'AMGA [133], le volet biomédical a été largement pris en compte car il n'existait pas de système permettant d'associer de façon sécurisée des métadonnées aux fichiers stockés sur grille. Il fallait alors proposer une politique de sécurité qui soit suffisamment robuste pour supporter un système distribué, naturellement plus exposé aux attaques.

En adoptant le modèle de sécurité de gLite, AMGA permet à tout utilisateur d'une VO de s'authentifier sur un serveur et de le consulter de façon totalement transparente. Cela est rendu possible grâce à l'intervention de VOMS, qui authentifie fortement un utilisateur. Cela permet aussi à un job sur la grille d'hériter des droits de son créateur tout au long de son exécution quel que soit le CE physiquement hôte du job.

Utilisations d'AMGA

Les usages d'AMGA sont maintenant nombreux dans les environnements de grille. De nombreux projets issus d'EGEE ont été séduits par les capacités d'AMGA et l'ont adopté. Parmi ces projets on peut en citer plusieurs :

- WISDOM¹ [134] qui propose un outil au déploiement de *data challenges* sur grille dans le domaine de la recherche pharmaceutique. L'ensemble du système d'information est stocké dans une base AMGA.
- MDM² [135] qui permet de connecter le système d'information de gLite avec les serveurs DICOM hospitaliers. Une image médicale inscrite peut alors être exposée au système d'information de la grille de façon transparente et sécurisée tout en restant physiquement dans le serveur DICOM. Les parties métadonnées de l'image DICOM sont stockées en lieu sûr dans un catalogue AMGA.

¹ Wide In-Silico Docking On Malaria

² Medical Data Manager

AMGA : sécurité, données médicales et lien avec VOMS

Intrinsèquement, AMGA a adopté un mécanisme d'*Access Control List* (ACL) qui permet, via une déclaration d'utilisateurs et de groupes d'associer à ces derniers des droits d'accès aux données. Ainsi, il est possible, pour chaque répertoire d'AMGA de restreindre ou d'autoriser un utilisateur ou un groupe d'utilisateurs à accéder aux données.

Dans la problématique de partage des données médicales évoquée ici, AMGA permet, en plus d'une authentification forte, d'autoriser ou non un utilisateur à accéder à un ensemble de données. Ainsi, il est possible, en séparant dans l'arborescence les données médicales des données personnelles, de répondre aux deux cas d'utilisation [2.4.2.3] et [2.4.2.4], des deux types de clients du réseau Sentinelle (associations de dépistage et épidémiologistes) évoqués dans le cahier des charges.

Grâce à la création des groupes `cancer_depistage` et `cancer_epidemio` dans VOMS, AMGA peut ainsi accorder des droits spécifiques aux répertoires en fonction des cas d'utilisation. En effet, contrairement aux associations de dépistage organisé, les épidémiologistes n'auront pas accès aux données confidentielles d'un patient.

3.2.3. Mise en œuvre des concepts de grille pour RSCA

La [Figure 33] représente à ce stade comment est constitué la grille et la VO dédiée à RSCA. Deux services sont déployés, au centre de l'infrastructure : VOMS, et AMGA. Ceux-ci fonctionnent de pair : VOMS définit la VO Sentinelle, les groupes et les rôles des utilisateurs et AMGA récupère cette information pour affecter les autorisations aux répertoires contenant les données.

Ainsi, un utilisateur appartenant au groupe *epidemio* n'aura pas accès aux données d'identité du patient.

La VO pourra ainsi accueillir de nouveaux composants afin de mettre en place l'échange de données à proprement parler. Chaque site relié à la VO Sentinelle, comme un laboratoire d'anatomie pathologique ou d'une structure de dépistage bénéficie alors de la même politique de sécurité, ce qui facilite le travail d'administration, objectif fonctionnel présenté en [2.4.2.2], maintenant centralisé autour de VOMS et repris dans AMGA.

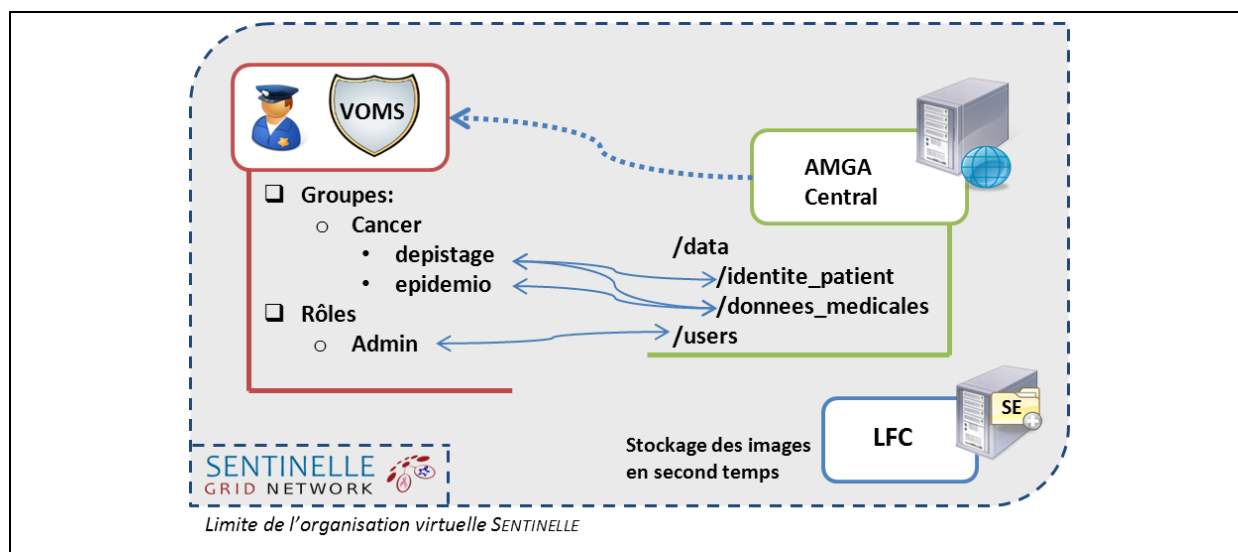


Figure 33 - Constitution de la partie administration de la VO Sentinelle

Le stockage des fichiers peut se faire via le LFC mais sera effectif dans un second temps, lorsque l'application évoluera vers le stockage d'images médicales, présenté comme objectif en [2.4.2.5].

3.3. L'ARCHITECTURE INFORMATIQUE POUR LE RESEAU SENTINELLE CANCER AUVERGNE

Les faiblesses de l'architecture client-serveur pour la gestion des données, présentées en [3.1.2.1], notamment sur la contrainte essentielle des acteurs d'anatomie pathologiques de ne pas sortir les données de leurs locaux imposent naturellement un système distribué pour l'accès aux données.

VOMS et AMGA constituent ainsi la base d'une Organisation Virtuelle « Sentinelle » dédiée à RSCA comprenant une hiérarchie de droits d'accès suivant les différents cas d'utilisation. Ce socle permet maintenant de s'intéresser aux développements qui sont encore nécessaires à la réalisation des objectifs fonctionnels. L'intégration de LFC permet ainsi de prévoir un système de stockage de fichiers lorsque le stockage d'images médicales sera nécessaire pour le projet.

3.3.1. Mise en œuvre de l'architecture

Partant de la VO Sentinelle, il reste encore quelques développements nécessaires à la mise en œuvre de l'architecture. Bien que VOMS et AMGA constituent les fondements de la sécurité et du mécanisme d'autorisation, il manque encore plusieurs éléments afin de remplir tous les objectifs du cahier des charges :

- L'intégration et la restitution des données : une solution technique devra être adoptée pour faciliter les entrées/sorties sur le réseau avec les fournisseurs des logiciels des acteurs du projet.
- L'authentification des utilisateurs : le cahier des charges [2.6.1.1] et la réglementation en vigueur [1.1.4] et [3.1.2.5] imposent une authentification forte par utilisation des CPS. Une méthodologie est à fournir pour coupler VOMS avec la certification CPS.
- Le composant logiciel principal : un serveur logiciel sera nécessaire pour réaliser toutes les cas d'utilisations définis dans le réseau.
- Le modèle de données : une structuration des données médicales doit être proposée, via un modèle générique capable d'intégrer des sources hétérogènes.

Le choix d'un langage ou protocole de communication à l'interface des entrées-sorties du réseau est aussi nécessaire pour l'interface avec les logiciels métiers des acteurs du projet.

3.3.1.1. Les services web et l'architecture orientée service

D'après le cahier des charges, une des principales contraintes du projet est de rendre un système capable d'interagir avec les sources de données anatomopathologiques, de les mettre en forme puis de les restituer aux clients du réseau, à savoir les associations de dépistage et les acteurs de santé publique.

Devant les grandes disparités entre les systèmes informatiques des différents acteurs du projet, les défis informatiques qui se heurtent à la réalisation de ce projet sont nombreux. De plus, les contraintes des anatomopathologistes imposent de laisser les données à l'endroit où elles ont été produites.

Les technologies informatiques actuelles s'adaptant au mieux à ces environnements fortement hétérogènes et distants, tout en permettant d'être accessibles à un ensemble de clients le plus simplement possible sont les services web.

Le principal avantage des services web réside dans leur capacité d'interopérabilité avec la quasi-exhaustivité des plates-formes et des langages informatiques. En effet, en utilisant des standards et protocoles ouverts de communication et de description des données, les échanges de données entre applications deviennent grandement facilités.

Une architecture orientée service, ou *Service Oriented Architecture (SOA)*, qui est une généralisation d'une architecture d'objets distribués, fonctionne sur le principe que chaque fournisseur de service peut être simultanément un client. Incités par les bénéfices potentiels en sécurité, disponibilité et facilité d'extension, de nombreuses SOAs ont été développées pour la création de dossiers médicaux partagés.

Point de vue technique

Un service web est une exposition sous forme logique d'un composant métier d'une application. Une architecture orientée service fournit alors une abstraction de toute la structure interne d'un logiciel, que ce soit le langage de programmation utilisé ou le SGBD sous-jacent. Les services sont alors faiblement couplés et peuvent être invoqués sans avoir aucune connaissance sur la façon dont ils sont intrinsèquement conçus.

La communication sous forme de messages standardisés que définit la norme SOAP¹ [136] via le langage XML (le plus souvent) facilite l'intégration par les logiciels tiers. De plus, chaque service est associé avec une description publique qui explique sa sémantique (à destination du programmeur) ainsi que son fonctionnement sous forme de métadonnées (à destination de l'outil de programmation) pour comprendre comment les méthodes doivent être invoquées. Le langage de description utilisé par les web-services est le format WSDL² [137].

Le référencement d'un service web se fait à travers UDDI³ [138], un annuaire XML de définition des services et de leur découverte, largement utilisé dans le cadre du commerce électronique.

Les services web constituent une exception notable au formalisme et conformisme habituel du comportement des différents langages de programmation lorsqu'il s'agit d'accéder à des objets distribués. En effet, un composant logiciel, interopérable avec une palette quasi-exhaustive de langages de programmation, auto-descriptif, utilisant des standards libres et reconnus de communication comme de présentation des données est un élément assez unique et innovant dans le monde logiciel.

¹ Simple Object Access Protocol

² Web Service Description Language

³ Universal Description Discovery and Integration

Ainsi, pour le réseau sentinelle, cette interopérabilité permettra, pour des clients identifiés comme les associations de dépistage ou les épidémiologistes ou potentiels (services de santé publique, ...) d'interfacer leur système d'information ou logiciel de gestion le plus simplement possible.

3.3.1.2. Authentification CPS et VOMS

L'utilisation des « smart cards » ou cartes à puce dans le milieu médical connaît un véritable essor [139]. La France a d'ailleurs été le premier pays à déployer à l'échelle du territoire, via son application Sesam-Vitale, une infrastructure utilisant ce type de matériel.

VOMS, présenté en [3.2.2.3] a été choisi comme le gestionnaire central de la sécurité sur la VO Sentinelle. Le cadre de sécurité proposé par l'infrastructure PKI [3.2.2.2] posera une base cryptographique suffisante au regard des recommandations ASIP [3.1.2.5] sur les dossiers médicaux partagés. Cependant, d'après ces mêmes recommandations [101, 103], il est indispensable d'utiliser l'infrastructure de Carte de Professionnel de Santé (CPS) pour toutes les phases d'authentification, de chiffrement de l'information et de signature électronique des actes médicaux.

La question principale est alors d'étudier la faisabilité d'une authentification par CPS sur une Organisation Virtuelle d'une grille utilisant VOMS.

Sur le site « éditeur » du GIP-CPS¹, toujours en activité malgré la fusion avec l'ASIP-Santé, est proposé un ensemble de documents techniques informant sur le fonctionnement des cartes CPS et de leur contenu.

D'après le GIP-CPS [140, 141], le contenu d'une carte CPS, présentée en [Figure 34] se compose de plusieurs éléments importants :

- les données personnelles professionnelles et techniques du professionnel de santé porteur de la carte, ainsi que les données nécessaires à la facturation pour l'assurance maladie ;
- les certificats électroniques, de signature et d'authentification, contenant des données techniques sur l'identité du porteur.

La partie nous intéressant ici concerne les certificats, ceux-ci [141], respectent la norme de cryptographie X.509 et l'infrastructure PKI. Ils sont théoriquement pleinement compatibles avec le modèle adopté sur les grilles propulsées par gLite et par VOMS.

La plus grande incertitude concerne l'accès au certificat par un client pour le fournir à VOMS. Pour cela, une étude matérielle des lecteurs de carte à puce [142] et des API² permettant d'accéder au contenu de ces cartes est nécessaire [143]. Par la suite, une évaluation de la compatibilité de la chaîne de certification fournie par la CPS est aussi de rigueur [144].

¹ <https://editeurs.gip-cps.fr>

² *Application Programming Interface* ou Interface de programmation

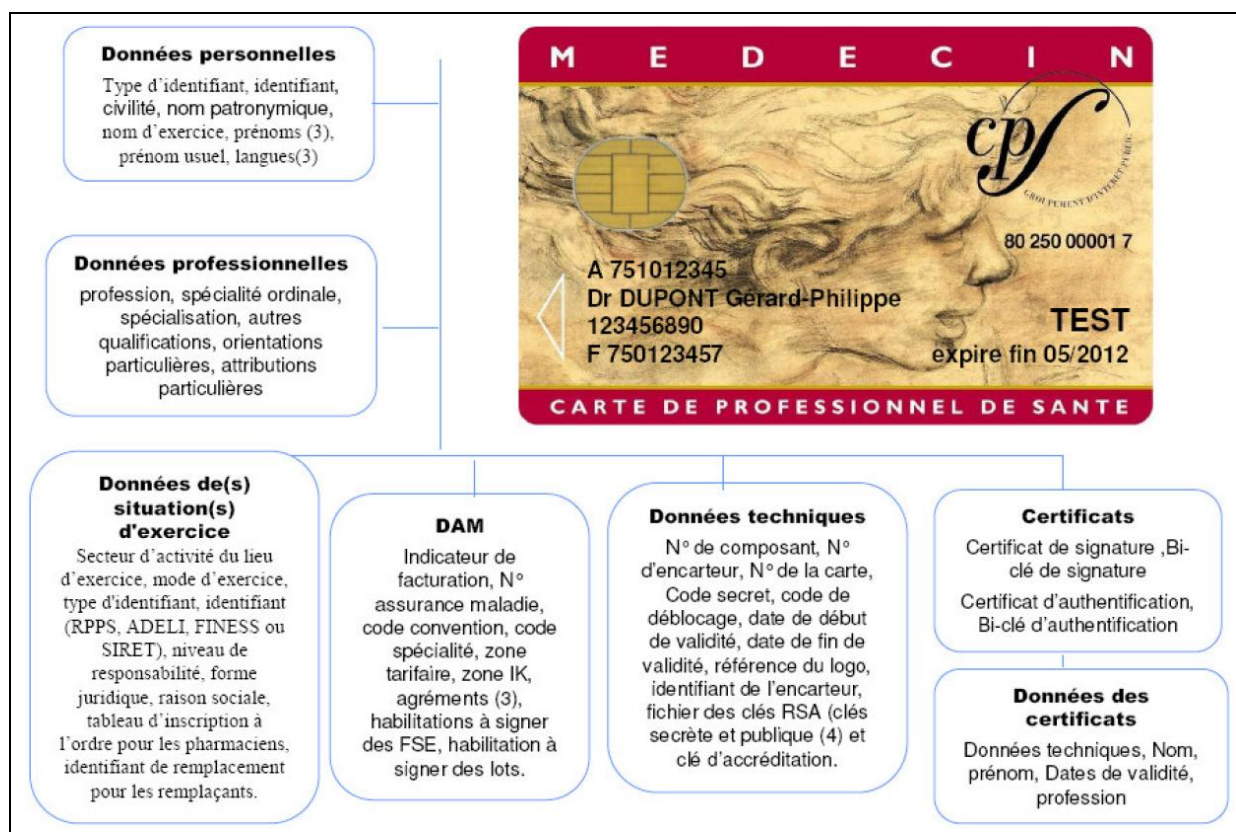


Figure 34- Contenu d'une carte CPS - Source Gip-CPS

PKCS

Fort heureusement, les cartes CPS utilisent les « *Public-Key Cryptography Standards* » (PKCS) et notamment le PKCS#11 qui régit l'utilisation des cartes à puces et du matériel de lecture pour le chiffrement et la certification.

PKCS [145] est un ensemble de 15 standards de cryptographie. Les plus courants sont PKCS7 et PKCS12 utilisés pour l'authentification personnelle en utilisant des certificats, comme par exemple les certificats grille.

La [Figure 35] montre le fonctionnement de PKCS#11 et les interactions entre le matériel et la couche applicative adapté à la CPS.

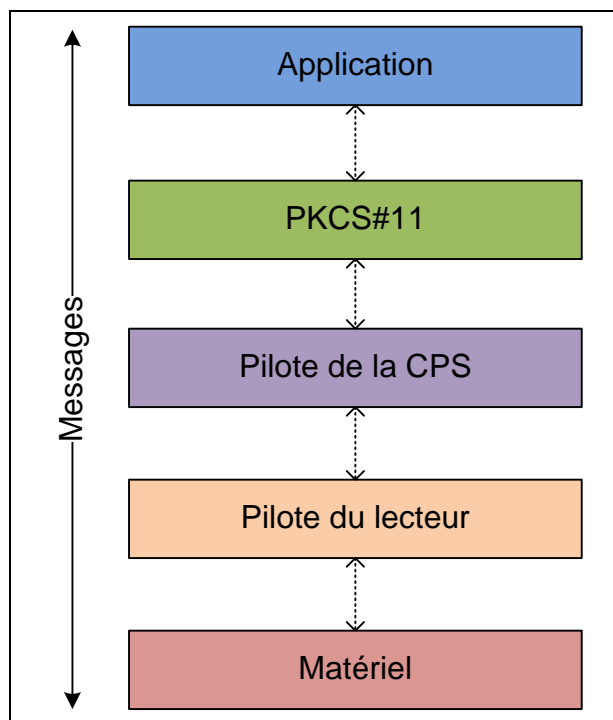


Figure 35 - Couches PKCS#11 pour CPS

Faisabilité

Compte tenu des points précédents, les éléments nécessaires à l'utilisation des certificats CPS sont réunis pour l'authentification sur une architecture de grille utilisant VOMS. Cependant, des développements sont à effectuer pour intégrer correctement tous ces éléments applicatifs au système d'authentification.

L'utilisation conjointe de PKI pour le modèle cryptographique, de VOMS pour l'authentification des utilisateurs, d'AMGA pour le contrôle d'accès et des CPS pour la certification des identités permettra de garantir le niveau de sécurité requis pour l'échange de données médicales en regard des contraintes de sécurité du cahier des charges [2.3.7.1] et des dispositions légales [1.1.4].

L'implémentation de la solution fait l'objet d'une attention particulière ultérieurement dans ce document, en [4.1.3]. Pour cela, un jeu de cartes de test accompagné de leur lecteur a été commandé auprès du GIP-CPS. Un abonnement à la section « éditeurs » a permis la récupération des éléments techniques (pilotes et API) ainsi que la documentation nécessaire à la mise en œuvre.

De cette manière tout membre voulant accéder à RSCA sera équipé d'un lecteur ainsi qu'une carte CPS personnelle.

3.3.1.3. Pandora Gateway

Le logiciel Pandora Gateway, qui sera abrégé par Gateway par la suite est une infrastructure développée par la société maat-Gknowledge [146] lors du projet Health-e-Child [89] qui peut être considéré comme une couche d'abstraction de la complexité des grilles pour des applications (bio)médicales. Ce projet vise à fournir une plateforme intégrée à destination des services de pédiatrie en Europe. Des hôpitaux de renom dont Necker de l'assistance publique des hôpitaux de Paris ou encore l'OPBG -Bambino Gesù Paediatric Hospital à Rome font partie des partenaires ayant en leur sein déployé la plateforme.

Le succès du projet Health-e-Child [147] a popularisé les technologies sous-jacentes qui ont été développées lors de son implémentation. La conception générique de la plate-forme la rend réutilisable pour le projet RSCA.

Fonctionnalités

Le logiciel réutilise un grand nombre de technologies et composants issus d'autres projets. En voulant créer une infrastructure charnière entre le monde médical et les grilles informatiques, le logiciel s'appuie naturellement sur des technologies de grille. Lors de la conception du logiciel, les technologies gLite et globus issus des projets de grille EGEE ont été les plus pertinents pour leur intégration. Afin de rendre plus simple l'intégration de systèmes de grille, l'utilisation de javaGAT¹ [148, 149] et SAGA² [150, 151], deux réécritures des commandes de grilles en Java, a été privilégiée.

Du point de vue du système d'information, leur choix s'est orienté sur AMGA qui présentait l'avantage d'être déjà pleinement compatible avec gLite et sa politique de sécurité.

Architecture

De manière générale, Pandora Gateway peut être considérée comme une SOA, donc pleinement compatible avec le standard de communication adopté en [3.3.1.1], à savoir les services web. Son architecture, comme le montre la [Figure 36] est scindée en composants principaux organisés en différentes couches.

A la base se situe une couche d'abstraction de l'intergiciel gLite, des technologies globus et du système de gestion de bases de données. A cette couche s'ajoute une partie « domain logic » qui fournit un ensemble de services génériques de bas niveau, qui seront réutilisés par la couche supérieure, « business logic » qui elle, fournit des services fortement spécifiques au domaine applicatif. Ce modèle permet un faible couplage entre les couches et minimise les développements nécessaires à la prise en charge d'un nouveau domaine. Seules les couches supérieures ont besoin d'être réécrites et adaptées.

¹ Java Grid Application Toolkit

² Simple API for Grid Applications

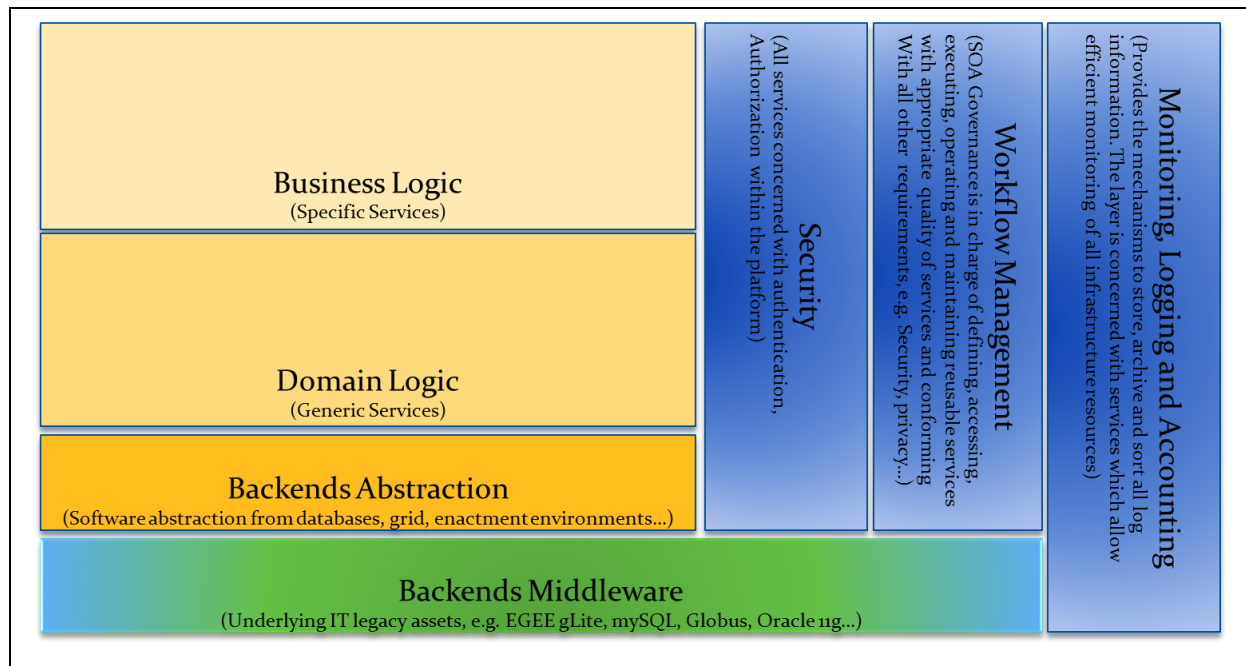


Figure 36 - Architecture globale de Pandora Gateway - Crédit maat-G

A ces couches s'ajoutent des composants essentiels de sécurité, de gestion des tâches sur grille et de suivi des activités de la plate-forme « logging and accounting ».

Fonctionnement, sécurité

La sécurité était un point important lors de la mise en œuvre de la plateforme. Les composants de gLite, l'infrastructure PKI [120] et le GSI [122] ne suffisent pas à garantir le niveau de sécurité requis pour le traitement de données nominatives à caractère médical mais peuvent former une clé des verrous logiciels. De plus, les web services, bien que pourvus de mécanismes de sécurité comme SAML¹ ou XML signature [152], peuvent souffrir d'un manque de sécurité. McGraw et al. [153], montrent que ce manque de sécurité est le plus souvent dû à un manque d'information et de compétences sur la sécurité des services web. Trop souvent axée réseau et communication (SSL/TLS), la sécurité des SOA doit être omniprésente dans toute la chaîne applicative. Globus a bien été conscient de ce manque et a donc renforcé la sécurité des services web pour la grille en adoptant un modèle d'autorisation calqué sur GSI et fourni un modèle de développement [154, 155]. Des « *security descriptors* » ont été ajoutés aux descriptions des services web. Ils sont de plusieurs natures :

- *Container Security Descriptor* ;
- *Service Security Descriptor* ;
- *Resource Security Descriptor* ;
- *Client Security Descriptor*.

Ils permettent de régler finement la sécurité, en utilisant les fichiers gridmap (ancêtre basique de VOMS qui reposait sur un fichier plat), avec une restriction des hôtes, ressources ou client. Pandora Gateway a proposé d'autres « *security descriptors* » personnalisés à ses exigences.

¹ Security Assertion Markup Language

C'est donc ce concept qui a été repris par la Gateway, en s'appuyant sur le système d'authentification et d'autorisation des grilles, qui s'appuient sur PKI pour fournir un degré supplémentaire de sécurité. La [Figure 37] montre comment cette sécurité influence l'accès aux services proposés par la Gateway. Un premier service, d'authentification^①, vérifie en plusieurs étapes la validité d'une requête d'un utilisateur. Tout d'abord il vérifie que l'authentification sur la grille fonctionne (via VOMS^②), ensuite il vérifie que l'utilisateur est bien habilité sur la Gateway en question, puis s'assure que les droits nécessaires sont présents. Ces trois conditions permettent alors de débloquent l'accès aux services métier de la Gateway^③, ainsi l'utilisateur peut les interroger normalement^④.

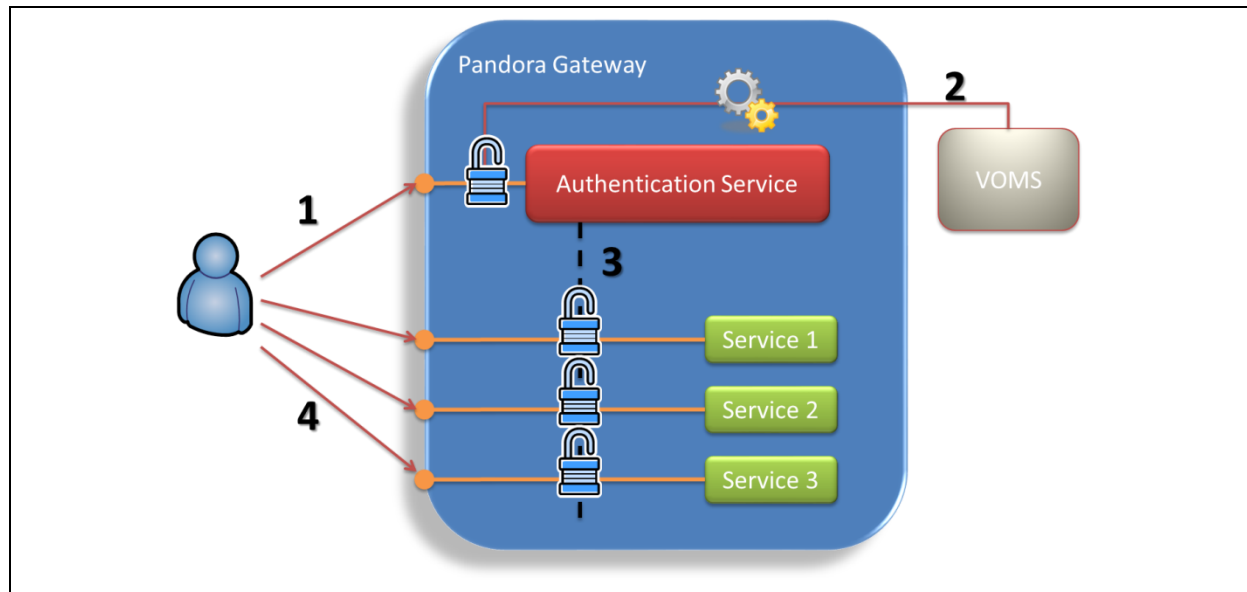


Figure 37 - Sécurité des services Gateway - Crédit maat-G

Le diagramme de séquence [Figure 38] récapitule les différentes étapes nécessaires au service d'authentification pour déverrouiller l'accès aux autres services. Une fois la connexion effectuée, avec le proxy grille créé, une pulsation régulière entre l'utilisateur et la Gateway signale sa présence. En cas d'arrêt de cette pulsation, l'accès aux services est refusé. Le proxy grille et les ressources sont détruits, ce qui empêche par conséquent toute utilisation de la grille. L'utilisateur doit alors recommencer la procédure d'authentification.

Communication intra-Gateway

Les échanges de données entre un réseau de Gateways sont primordiaux pour la cohérence d'une infrastructure de ce type. Ainsi, les concepteurs de l'application l'ont pourvu de primitives dédiées à l'échange de données. Celles-ci sont de deux types :

- la première repose sur la réplication intégrée à AMGA. Celle-ci diffuse le système d'information de l'ensemble du réseau, maintenant ainsi une information sur son état ;
- la deuxième s'appuie sur les services, un utilisateur authentifié sur une Gateway pourra voir ses droits délégués dans une autre instance distante appartenant à la même VO sur la grille.

Ces deux éléments permettent ainsi de réellement considérer une Gateway comme l'élément central d'un nœud de la VO Sentinelle, capable de gérer à la fois la sécurité et les entrées-sorties sur les données.

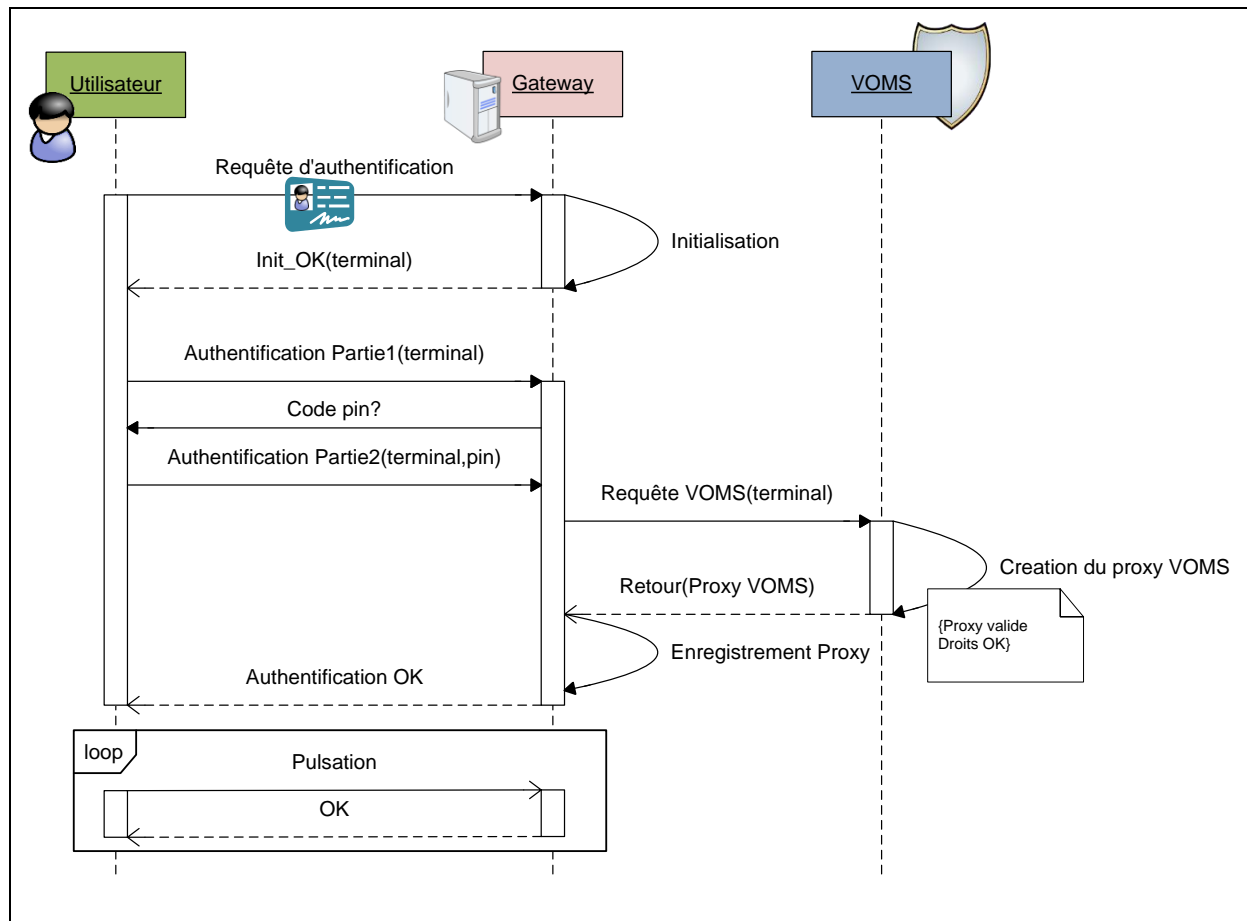


Figure 38 - Diagramme de séquence - Service d'authentification Gateway

3.3.1.4. Choix du modèle de données : Integrated Case Data

Au sein de la Gateway a été développé le modèle « *Integrated Case Data* » (ICD). Ce modèle de représentation des données médicales [89], utilise plusieurs ontologies dont UMLS [156], Galen [157] ou GeneOntology [158] pour effectuer une description sémantique des données médicales.

La conception générique d'ICD en fait un candidat parfaitement adapté au contexte de RSCA qui consiste à intégrer diverses sources de données et les rendre interopérables pour des requêtes statistiques globales.

Cette représentation, en plus d'une annotation sémantique qui permet de connaître le véritable contenu des données, facilite l'intégration et la restitution de celles-ci, voir [Figure 39].

Intégration de données

Le modèle ICD interne étant défini pour le domaine étudié, il est alors possible d'intégrer des données provenant de sources diverses avec une représentation différente. Pour cela il suffit alors, si la source est sémantiquement annotée, d'aligner les ontologies puis de créer le processus de transformation. Le cas échéant, ce processus devra être fait ad-hoc, c'est-à-dire mettre en correspondance les champs des sources de données et préciser les transformations nécessaires. Compte tenu de la faible disponibilité de données annotées dans le cadre de RSCA, ce dernier processus sera considéré.

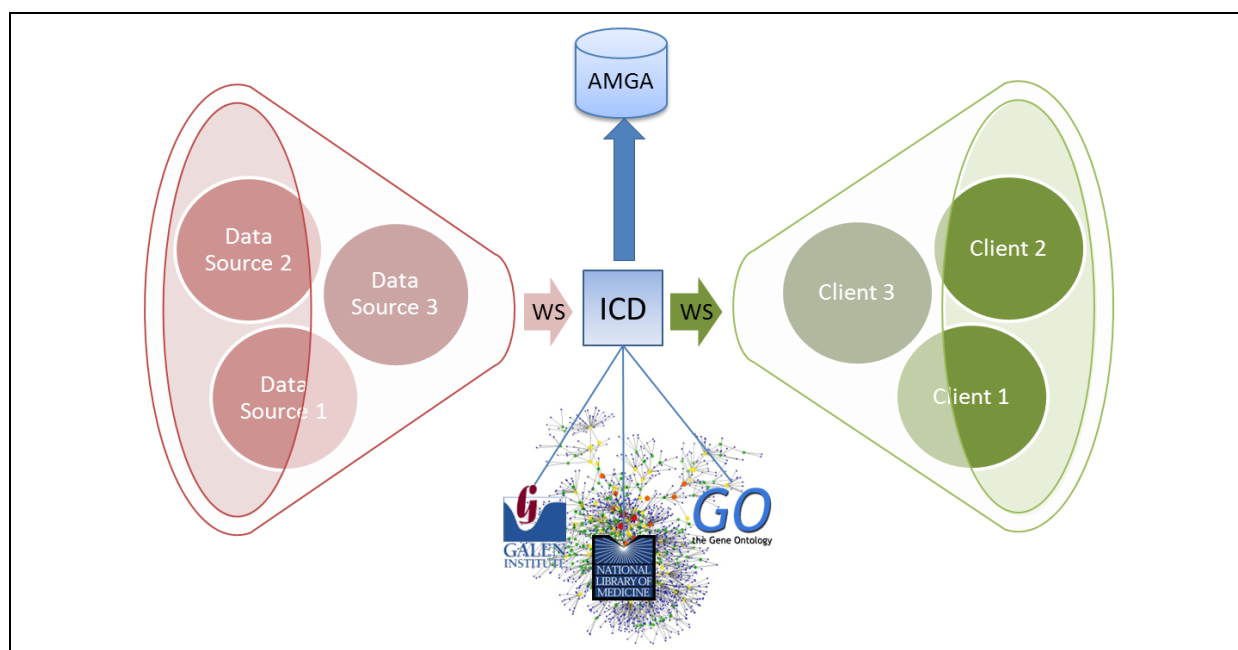


Figure 39 - Système « Integrated Case Data »

Restitution de données

Une fois les données intégrées dans le système d'information de la Gateway, il est possible de créer de multiples schémas d'export en fonction de la demande utilisateur. Ainsi l'utilisateur peut choisir les champs à exporter ainsi que leur format. De plus, afin de répondre aux besoins des épidémiologistes certains calculs et regroupements sont possibles pour faciliter le prétraitement des données.

Concepts internes

Le schéma de données développé dans ICD a été conçu de manière la plus générique possible pour représenter l'information médicale. Ainsi, quatre principaux concepts ont été introduits :

- *Patient* : représente l'information sur un patient, son identité.
- *Visit* : indique l'objet de la visite d'un patient ainsi que sa date, nécessaire pour établir le contexte d'un examen médical.
- *Medical Event* : Indique l'examen subi lors de la visite du patient (auscultation, imagerie, acte chirurgical, ...).
- *Clinical Variable* : Ensemble de données médicales acquises lors d'un examen: image médicale, signes vitaux : tension artérielle, pouls, etc. Ces variables ont le plus souvent un ensemble de métadonnées les caractérisant, comme l'unité, leur signification ou encore les métadonnées des images médicales DICOM [1].

Les relations entre ces différents concepts est de type 1-à-N, c'est-à-dire qu'un patient peut effectuer plusieurs visites qui concernent plusieurs examens qui ont permis d'acquérir plusieurs variables cliniques. Dans le cas de RSCA, il n'est pas nécessaire de procéder à une annotation sémantique car les sources de données possèdent le même format dans les bases distribuées entre les différents services de pathologie.

La séparation des concepts permet aussi d'introduire des droits spécifiques à chacun d'entre eux. Comme précisé dans le cahier des charges et en partie [3.2.2.8], suivant le cas d'utilisation, certains utilisateurs n'auront pas accès aux informations concernant l'identité du patient.

Intégration d'une source de données à ICD

Les différents concepts internes à ICD cités précédemment permettent de représenter la plus grande partie des informations médicales d'un patient de façon hiérarchique. Partant des sources de données anatomo-pathologiques, il faut alors convertir les informations au format ICD en séparant les concepts.

La [Figure 40] présente les différentes étapes : premièrement, à partir du modèle source à intégrer, il faut spécifier la représentation des données et des métadonnées en correspondance au modèle abstrait ICD. Ensuite, les différents concepts sont identifiés au sein de la source pour créer un importateur qui va intégrer les données à la Gateway et ainsi créer la base de données du réseau sentinelle. Un exemple de mise en application pour les données anatomopathologiques du cabinet SIPATH est présenté ultérieurement en [3.4.1.1].

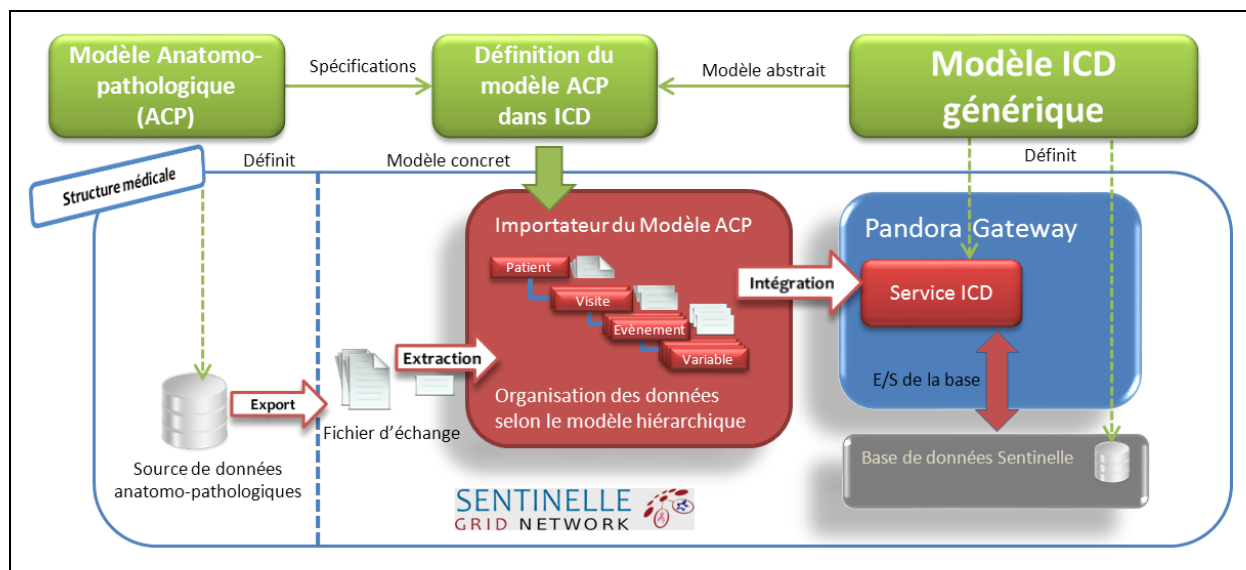


Figure 40 - Méthodologie d'import des données dans ICD

3.3.2. L'architecture choisie

Les différents choix technologiques, techniques et logiciel effectués permettent maintenant d'établir l'architecture de RSCA. Ainsi, l'infrastructure est constituée d'un noyau central chargé d'orchestrer la sécurité, de gérer les utilisateurs, leurs droits et de définir les autorisations au données par le biais du modèle de données ICD implémenté dans AMGA.

Par la suite, les différents nœuds pourront être intégrés de façon dynamique à la grille en fonction des sites à équiper. La VO sera dimensionnée proportionnellement au nombre de machines connectées.

3.3.2.1. Architecture complète

La [Figure 41] montre l'architecture du réseau, composée du noyau dur de la grille au centre et d'entités clients/fournisseurs de données gravitant autour. Les différents liens entre les sites schématisent une communication entre ceux-ci. Ils peuvent être de plusieurs natures en fonction du cas d'utilisation du réseau mais toujours sous contrôle de l'autorité de sécurité centrale sur la grille. Celle-ci pourra ainsi, en fonction de la nature de la requête l'autoriser ou non par rapport à la

politique d'accès. Elle sera aussi en mesure de tracer toute l'utilisation du réseau en rapport avec les contraintes fixées par la CNIL [1.1.4.2].

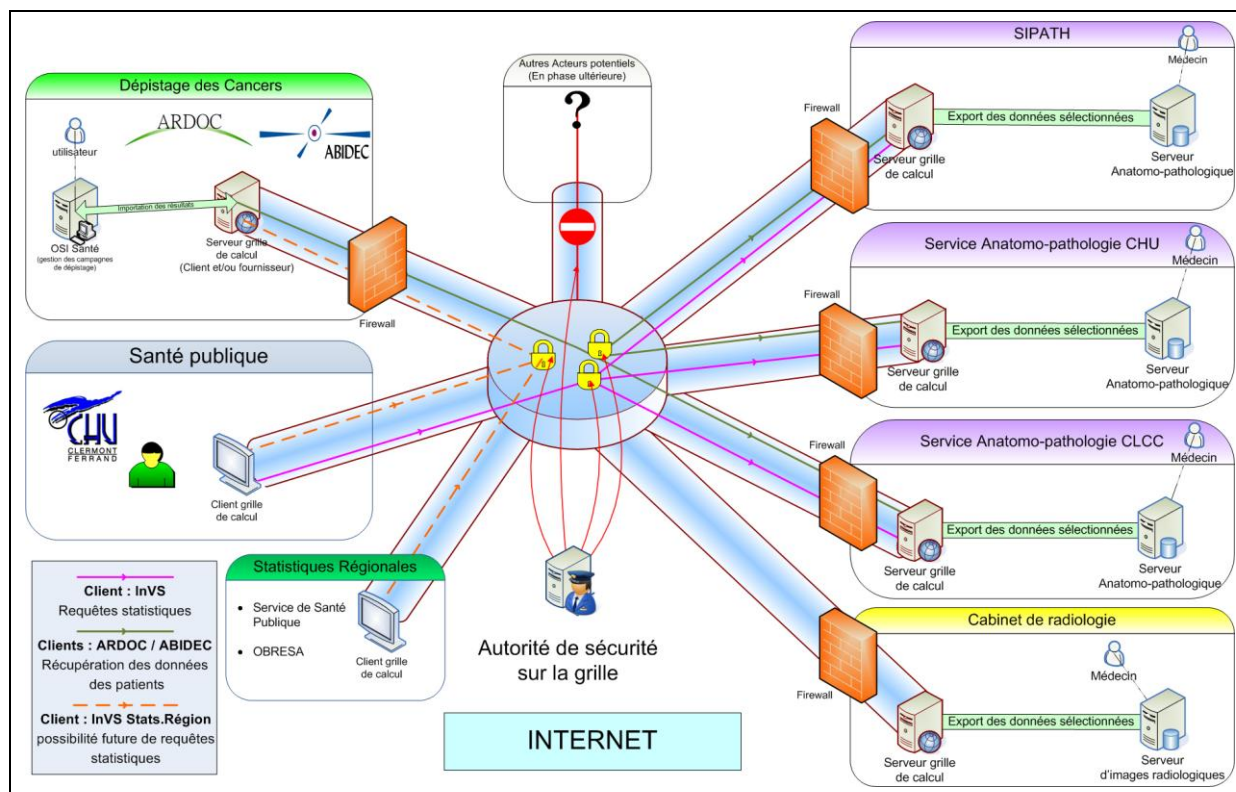


Figure 41 - Architecture complète du réseau

3.3.2.2. Composition du noyau central

Le noyau central de la grille, représenté sur la [Figure 41] par l'Autorité de sécurité est constitué principalement des services de base de gestion de la VO, à savoir VOMS et AMGA en mode maître pour la réplication. Il définit le modèle de données et les autorisations qui seront diffusées sur toute la VO.

Un serveur Web est aussi hébergé dans ce noyau central, à des fins de dissémination de communication. Il est aussi l'hôte de l'interface d'administration de VOMS permettant de définir les utilisateurs et leurs droits en fonction des groupes et des rôles qu'ils bénéficient. Par ailleurs, le serveur web est nécessaire pour fournir une interface aux requêtes statistiques des épidémiologistes.

L'autorité de sécurité de RSCA est présentée au cœur de ce dispositif car c'est elle qui gèrera à ce niveau l'exploitation complète du réseau sentinelle.

3.3.2.3. Composition des nœuds périphériques

La [Figure 41] représente un ensemble de nœuds interconnectés via le noyau central. Celui-ci héberge le serveur AMGA maître qui réplique tout le système d'information du réseau : renseignements sur les sites connectés et définition des utilisateurs. Il contient aussi le squelette du schéma ICD qu'implémenteront les sites reliés.

Les services d'anatomopathologie hébergent un serveur physique contenant une instance de la Gateway reliée via la chaîne de certification au noyau central et donc de facto, aux autres sites.

L'instance d'AMGA interne sera synchronisée (en esclave) avec le système d'information situé dans le noyau central. Celle-ci aura aussi connaissance du modèle ICD qu'elle implémentera et fournira en données anatomopathologiques. Ainsi tous les sites de pathologie partageront le même schéma fondé sur ICD, ce qui permettra, via les fonctionnalités de collaboration internes à AMGA [133], de répartir les requêtes de façon totalement transparente, comme montré en [Figure 42].

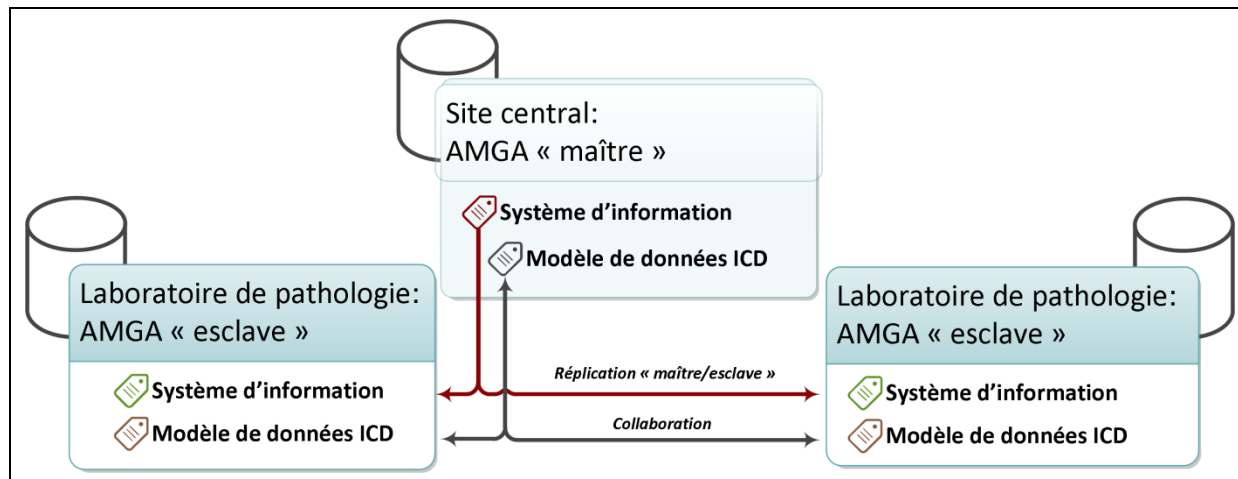


Figure 42 - Système de réplication et de collaboration dans AMGA

Les associations de dépistage seront équipées de la même façon sauf que dans un premier temps elles ne fourniront pas de données. La présence d'un serveur physique est tout de même nécessaire pour l'interface avec les logiciels métiers, préférable dans le réseau local.

En ce qui concerne les structures de santé publique, celles-ci n'ont pas besoin d'une installation d'un serveur dédié. Elles pourront alors, via les certificats CPS directement se connecter au noyau central pour effectuer leurs requêtes statistique au moyen d'un client léger (navigateur).

3.3.2.4. Analyse

La mise en application du réseau sentinelle s'appuie donc sur une architecture de grille privée, dérivée des composants gLite avec comme composant principal la Gateway. Les points les plus cruciaux évoqués dans le cahier des charges étant la sécurité des données, présentée comme objectif technique en [2.3.7.2] et en [2.6.1] du cahier des charges et en respect de l'objectif fonctionnel principal présenté en [2.3.7.1] de ne pas sortir les données des structures propriétaires, l'architecture de grille décentralisée permet de répondre à ces contraintes.

Premièrement le réseau propose à chaque structure de se raccorder au moyen de l'installation d'une Gateway en son sein, ce qui évite l'export des données à l'extérieur des sites de production. Ensuite, la sécurité des communications est assurée par l'utilisation systématique de PKI, rendant ainsi sûr l'échange de données. L'authentification des utilisateurs, uniquement possible par le moyen d'une carte CPS permet, en plus de s'assurer de l'identité des personnes, au moyen de l'utilisation conjointe de VOMS et d'AMGA, de proposer un système robuste d'autorisation.

D'un point de vue fonctionnel, comme présenté en partie [2.3.7.1] du cahier des charges, l'architecture ainsi créée répond, avec l'exposition des données via des web services au besoin d'interopérabilité des clients des données. L'utilisation du schéma ICD et son interrogation via des web services permettent de proposer une interface commune aux clients pour interroger la base du réseau sentinelle.

3.4. MISE EN APPLICATION POUR LE DEPISTAGE ORGANISE DES CANCERS DU SEIN ET DU COLON

Dans le cas d'utilisation qui consiste à mettre en relation les structures de pathologie et les structures anatomopathologiques la quasi-totalité des éléments nécessaires ont été mis en œuvre : une architecture distribuée respectant les objectifs et contraintes du cahier des charges est proposée. D'un point de vue matériel, un serveur de grille est installé au sein de chaque structure de pathologie puis relié à la fois à leur réseau local et au réseau extranet afin d'être accessible depuis l'extérieur. Il en est de même pour les associations de dépistage afin de le permettre d'accéder de façon sécurisée aux données. L'utilisation des web-services facilitera l'intégration des données dans les logiciels métier

3.4.1. *La structuration des données médicales*

Il reste cependant encore une dernière étape pour l'application à la problématique cancer, c'est-à-dire la structuration des données d'anatomie pathologique pour l'intégration dans le réseau et leur restitution pour les clients des données, à savoir les structures de dépistage organisé des cancers et les services de santé publique.

Le but est alors d'adapter le modèle ICD, présenté en [3.3.1.4] avec la structure des données anatomopathologiques au sein de la Gateway. Le cahier des charges du projet fait déjà état des informations à récupérer en [2.5] et en [2.9.1].

3.4.1.1. *Modèle anatomo-pathologique pour les structures médicales en Auvergne*

Les structures pathologiques participant au réseau ont été en mesure de nous proposer, via des comptes rendus médicaux anonymes, un modèle de données les concernant.

En [Annexe 1] se trouve un exemple de compte rendu pour une tumeur mammaire, source du modèle à adopter pour l'intégration des données dans le réseau. Il est constitué de 6 parties différentes :

- le médecin prescripteur, ses coordonnées ;
- le patient, son état civil et adresse complète ;
- les informations sur le prélèvement à analyser, avec les différentes dates ;
- le médecin préleveur, ayant effectué l'acte chirurgical sur le patient ;
- les informations sur l'examen anatomo-pathologique ;
- la conclusion médicale, sous forme textuelle ;
- la codification ADICAP.

Le stockage informatique de ces données est quant à lui légèrement différent, la récupération des données s'est faite en collaboration avec l'éditeur informatique, présenté en [Annexe 1]. Une partie au format XML est structurée, contenant la majorité des données administratives : médecins, patients, adresses, numéros d'identification du dossier. La seule information médicale présente dans ce dossier étant la codification Adicap, présentée en [1.2.1.2]. Le restant des données médicales se trouve alors dans un fichier au format .doc non structuré.

Modèle de données

Vu que tous les logiciels de gestion des structures d'anatomie pathologiques d'Auvergne sont issus du même éditeur, le schéma de données sera unique sur l'ensemble du réseau.

Le modèle de données, à partir des informations collectées est alors établi, comme montré en [Figure 43]. Le nom des champs est issu du logiciel métier anatomopathologique, la valeur de ce champ est précisée dans la deuxième colonne. Le lien avec ICD est représenté par un code couleur en légende de la figure. Ainsi, parmi les quatre concepts Patient, Visit, Medical Event et Clinical Variable, seuls les trois derniers seront disponibles aux épidémiologistes.

Nom des champs	Valeur	Concepts ICD :	
NURES ID	Identifiant dossier interne à la base de données	<div><div></div> Patient</div> <div><div></div> Visit</div> <div><div></div> Medical Event</div> <div><div></div> Clinical Variable</div>	
NUDTEXT	Identifiant de l'examen (public)		
DATPREL	Date ou le prélèvement a été effectué		
DATENREG	Date ou le prélèvement a été reçu au labo		
NUPAT	Numéro du patient		
NOMPAT	Nom patronymique	<div><div></div> Patient</div> <div><div></div> Visit</div> <div><div></div> Medical Event</div> <div><div></div> Clinical Variable</div>	
PRENOM	Prénom		
ADRESSE1	Adresse du patient		
ADRESSE2	Complément d'adresse 2		
ADRESSE3	Complément d'adresse 3		
CODPOSTAL	Code postal du patient		
VILLE	Ville		
PAYS	Pays		
SEXE	Sexe		
DATNAISSANCE	Date de naissance		
NOMLEC	Médecin lecteur		<div><div></div> Patient</div> <div><div></div> Visit</div> <div><div></div> Medical Event</div> <div><div></div> Clinical Variable</div>
INSEELEC	Code INSEE du médecin lecteur		
NOMLEC2	Médecin lecteur 2		
INSEELEC2	Code INSEE du 2ème médecin lecteur		
NOMMED	Nom de médecin (prescripteur)		
INSEEMED	Code INSEE du médecin prescripteur	<div><div></div> Patient</div> <div><div></div> Visit</div> <div><div></div> Medical Event</div> <div><div></div> Clinical Variable</div>	
NURES	Numéro dossier, interne à la BDD		
DATVALIDATION	Date de validation	<div><div></div> Patient</div> <div><div></div> Visit</div> <div><div></div> Medical Event</div> <div><div></div> Clinical Variable</div>	
MODPREL	INFORMATIONS ADICAP		
TYPTECH			
ORGANE			
LESION			
DATCODAGE	Date de finalisation	<div><div></div> Patient</div> <div><div></div> Visit</div> <div><div></div> Medical Event</div> <div><div></div> Clinical Variable</div>	
NOMORIG	Origine du prélèvement		
RESULTATCCL	Lien vers le fichier	<div><div></div> Patient</div> <div><div></div> Visit</div> <div><div></div> Medical Event</div> <div><div></div> Clinical Variable</div>	

Figure 43 - Modèle de données - Compte rendu anatomo-pathologique

Dichotomie traits d'identification - données médicales

La séparation offerte par ICD entre les concepts Patient et Visit/Clinical Exam permet de garantir davantage la confidentialité des patients, les traits d'identification et l'examen seront séparés au

niveau base de données. Cette mesure peut en plus s'accommoder d'une définition des autorisations au niveau d'AMGA en accord avec VOMS en respect des différents cas d'utilisation [2.4.2.3] et [2.4.2.4] du cahier des charges. Ainsi, un utilisateur en épidémiologie du réseau n'aura pas accès aux données d'identification mais seulement au compte rendu médical.

Cette dichotomie a aussi son importance au niveau de la couche d'identification des personnes, présentée ultérieurement en [4.3], en évitant ainsi de transporter de l'information médicale avec les informations du patient lors de la phase d'ajout des données au réseau.

3.4.2. L'interfaçage avec les bases de données de pathologie

Avec le désir initial de créer un réseau réactif, permettant d'accéder aux données en temps très court, il fallait doter le réseau d'une interface avec les bases de données ou logiciels métier qui soient à la hauteur de cette ambition.

3.4.2.1. Solution pour le cabinet SIPATH

D'un point de vue logiciel, la solution informatique adoptée par le cabinet SIPATH est fournie par la société Infologic [159] et se nomme Diamic [160]. Ce logiciel est bâti sur le modèle purement serveur (données+applicatif); les utilisateurs se connectent à distance sur celui-ci pour l'utiliser. La base métier est alors centralisée sur le serveur de données.

Du fait de la complexité de la solution DIAMIC et des clauses de contrat qui lient la société éditrice et le cabinet de pathologie, il n'est pas possible d'aller directement chercher l'information dans les bases.

Point de vue réseau

Afin de garantir que les données restent dans le lieu où elles sont produites et de ne pas interférer avec la machine hébergeant DIAMIC il est nécessaire :

- d'isoler le serveur dans la structure médicale ;
- régler le pare-feu pour n'autoriser le serveur qu'en réception de données « mode push » avec le réseau local ;
- le firewall est configuré pour laisser entrer une liste prédéfinie d'hôtes sur internet (autres nœuds de la VO Sentinelle autorisés).

Point de vue données

Après discussion avec l'éditeur logiciel, la solution la plus avantageuse, en termes de développements, de rapidité d'exécution et de coût de déploiement est de communiquer sous forme d'un export « fichier » qui serait ensuite retraité au sein du réseau.

Hormis les avantages en termes de coûts, cette solution est aussi très faiblement intrusive. En effet, le fichier peut être « poussé » vers le serveur, limitant les communications à un lien unilatéral Serveur-métier -> Serveur-réseau sentinelle. De plus, la politique de sécurité réseau peut facilement s'adapter avec ce serveur car sa vocation est d'être uniquement dédiée à la réception de fichiers.

Cette solution, qui découple au maximum le lien entre le serveur et la source de données, permet aussi de faciliter l'entrée de nouveaux sites dans le réseau, avec des logiciels métier

différents comme TD-Synergy de Technidata qui fournissent d'autres laboratoires de pathologie en France.

Ainsi, une prestation de service auprès d'Infologic a permis de mettre en place cet export de données, « poussé » vers le serveur relié au réseau, comme montré en [Figure 44].

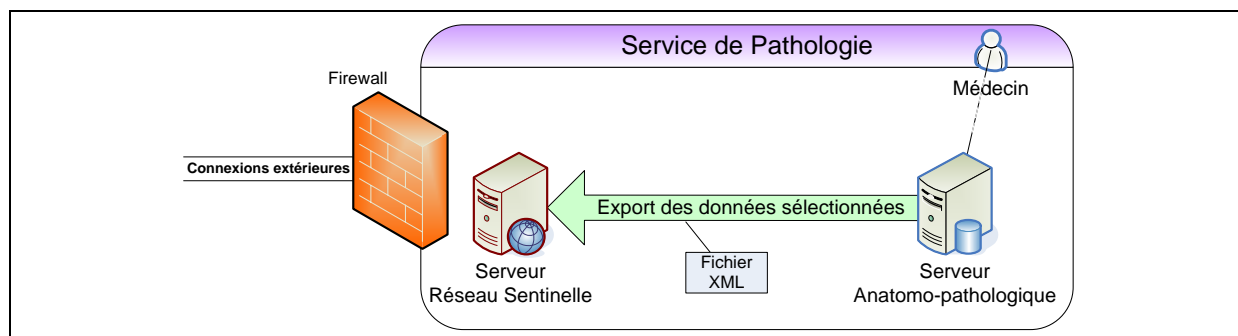


Figure 44 - Interfaçage avec les Logiciels Existants

Le format de fichier intermédiaire s'est porté sur XML, qui a l'avantage de structurer à minima l'information et qui ne demandait pas d'effort particulier de la part de l'éditeur.

L'intégration dans le modèle ICD [3.3.1.4] est aussi facilitée avec un document structuré en XML. Cependant, le document additionnel au format .doc ne permet pas d'être traité de la même façon et ne peut qu'être stocké que sous forme de pièce-jointe au modèle de données et ajoutée en tant que variable clinique.

Bien que cette solution n'offre pas un système « temps réel » entre le réseau sentinelle et les services de pathologie, contrairement à ce qui a été présenté au Chapitre 1, le fait de « pousser » les données vers le réseau sentinelle de façon périodique satisfait les contraintes du cahier des charges. En concertation avec les clients du réseau sentinelle, que ce soit les associations de dépistage ou les organismes de santé publique, un envoi hebdomadaire est largement suffisant pour satisfaire leurs contraintes. A fortiori, un simple réglage permettrait un envoi journalier des données si la demande se faisait sentir.

3.4.3. L'interfaçage avec les systèmes d'information métier

De l'autre côté du réseau se trouvent les organismes qui souhaitent récupérer les données, que ce soit les structures de dépistage ou les services de santé publique. De la même manière qu'il est nécessaire d'interfacer avec les bases anatomo-pathologiques, les données doivent être présentées de façon à être retraitées par les clients du réseau.

En fonction du volume de données à échanger entre le client et le réseau, il est possible de procéder de deux façons différentes :

- soit en interfaçant directement le poste client avec un serveur intégré au réseau, en installant la partie client du logiciel Gateway ;
- soit en ajoutant un serveur Gateway directement au sein de la structure à équiper, qui deviendra alors un élément à part entière du réseau.

3.4.3.1. Solution pour les structures de dépistage

Le cahier des charges, en [2.4.2.4], précise que les structures de dépistage pourraient devenir aussi fournisseurs de données statistiques. Ainsi, la deuxième manière de procéder est de mise dans ce cas: un serveur est installé à l'intérieur de la structure et fait partie intégrante du réseau, comme le montre [Figure 45].

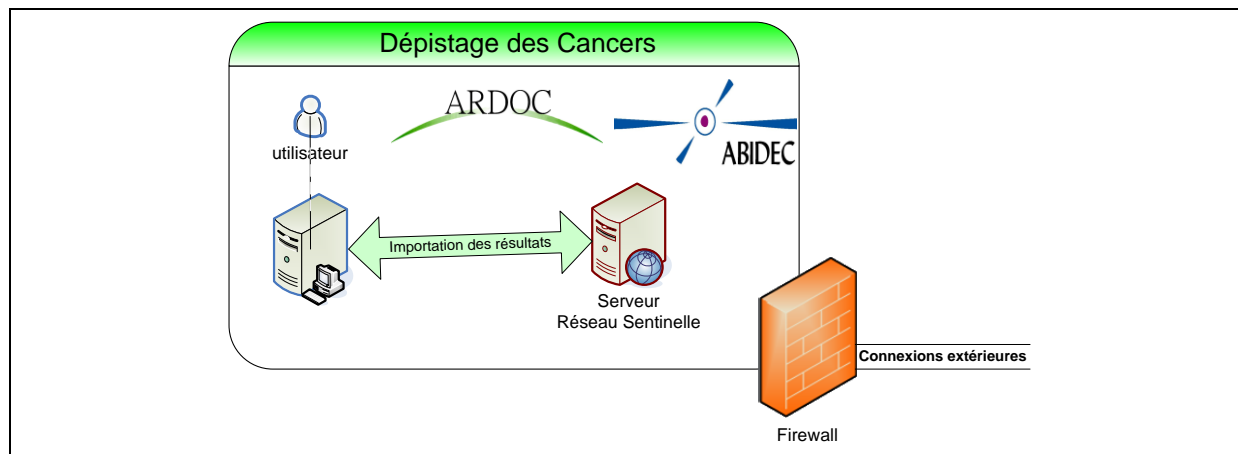


Figure 45 - Interfaçage avec les logiciels métier – Point de vue dépistage

L'import des données dans le système d'information de la structure de dépistage reste à la charge de leur éditeur logiciel. Ainsi, le système existant pourra interroger le service ICD local situé dans la Gateway pour récupérer les données. Compte tenu de l'universalité des services web, la majeure partie des clients pourront s'interfacer de façon bilatérale avec la Gateway.

3.4.3.2. Solution pour la santé publique

Les structures de santé publique ne seront jamais fournisseurs de données donc un simple poste client permettant de récupérer l'information sera nécessaire, comme montré en [Figure 46].

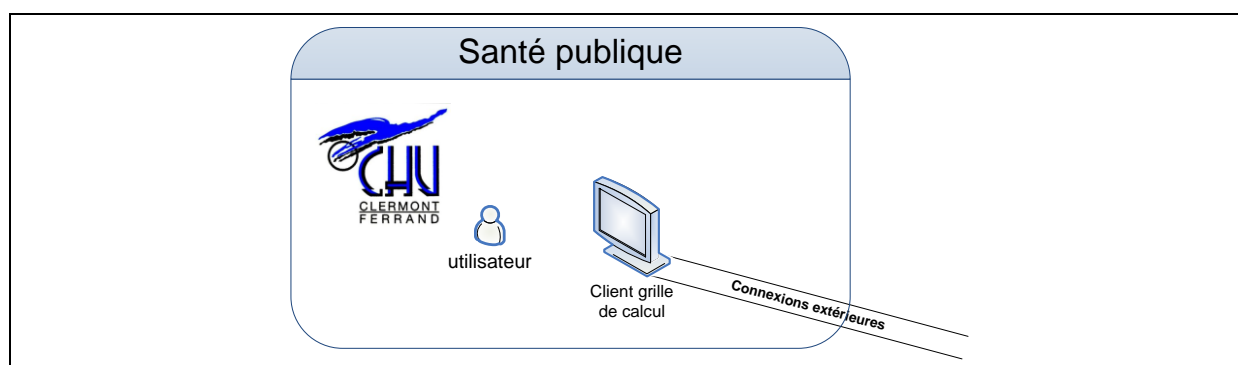


Figure 46 - Interfaçage du réseau sentinelle avec la Santé publique

Les clients devront s'authentifier avec leur carte CPS directement à la Gateway Centrale qui leur donnera accès, en fonction des autorisations, aux données du réseau sentinelle. Une interface pourra être fournie à la demande pour simplifier les requêtes dans le modèle AMGA.

3.4.4. Bilan de mise en place du réseau

L'architecture de RSCA est maintenant déployée, ainsi, comme indiqué précédemment :

- le noyau central du réseau est mis en service. Il est constitué d'un ensemble de serveurs sous la responsabilité de Maat-G. Ceux-ci hébergent VOMS ainsi que les éléments centraux de la VO, le serveur AMGA maître, le serveur BDII, le LFC et un SE. Un site a été équipé dans le réseau des grilles « Auvergrid » pour le développement, les tests et l'implémentation des services ;
- le cabinet SIPATH a été équipé : l'intégration des données a été effectuée et le modèle de données est défini dans ICD. Le site est maintenant complètement et automatiquement relié à RSCA. Depuis la mise en place de l'export automatisé des données, effectif au 01/01/2010, environ 1000 comptes rendus anatomopathologiques sur le cancer du sein sont expédiés par an, soit entre 10 et 30 par semaine. L'ARDOC doit récupérer environ 600 dossiers par an. Les autres laboratoires de pathologie fonctionnent avec le même logiciel et pourraient ainsi être raccordés sur le même schéma ;
- les deux associations de dépistage organisé des cancers d'Auvergne sont aussi reliées au réseau, et peuvent théoriquement effectuer des requêtes sur les données. En ce qui concerne l'ARDOC, environ 70000 dossiers patient peut être reliés au réseau ;
- les structures de santé publique, et plus particulièrement le service de santé publique du CHU de Clermont-Ferrand peut se connecter au travers de la Gateway dédiée située dans le réseau Auvergrid. Les requêtes sur les données sont gérées directement par AMGA. Un simple poste client doté d'un lecteur CPS est nécessaire.

Cependant, en l'absence d'une autorisation par la CNIL, il n'a pas été possible de procéder à la mise en œuvre effective de transfert de données. Ainsi, les échanges ne sont pas encore effectifs entre les sites.

CONCLUSION

La mise en place de l'architecture du réseau sentinelle est maintenant établie. Elle prend en considération les différents objectifs et contraintes du cahier des charges et propose une solution adaptée à la spécificité de RSCA.

Les technologies des grilles informatiques sont analysées et réutilisées en fonction de leurs atouts. Ainsi, VOMS et AMGA s'imposent naturellement comme clé de voûte de la sécurité, de l'authentification et de l'autorisation. L'utilisation de la Gateway a permis de faciliter grandement l'intégration des solutions de grille, en réutilisant et incorporant les technologies issues de globus et gLite dans une solution unifiée. Ainsi, un réseau constitué d'un ensemble de Gateways forme une grille informatique dédiée et spécialisée pour la problématique posée par le réseau sentinelle.

La gestion des données et la création d'un modèle conforme aux spécifications d'ICD propose une séparation entre l'identification du patient et données le concernant. Cette dichotomie permet de bien diviser les cas d'utilisation des associations de dépistage et des structures de santé publique. La création du modèle unique dans AMGA au sein des bases distribuées permet de rendre transparent les requêtes sur l'ensemble des sources simultanément, que ce soit pour une demande d'un compte rendu ou pour une analyse statistique.

Une mise en application complète est effectuée pour l'application cancer, en proposant une méthodologie d'import des données anatomopathologiques pour un laboratoire, reproductible sur les autres sites à équiper. En aval, l'interfaçage pour les associations de dépistage organisé et les structures de santé publique est proposé en fonction des cas d'utilisation et des contraintes fixées par le cahier des charges.

Cependant, certains éléments présents dans le cahier des charges n'ont pas tout à fait abouti ou n'ont pas pu être traités :

- la gestion des images médicales, sur le même modèle que les données de pathologie n'a pas encore été intégrée, bien que les outils préalables à la gestion des fichiers aient été mis en place dans la VO Sentinelle en [3.2.2.7] ;
- la mise en œuvre effective du réseau n'a pas pu se faire sur l'ensemble des sites prévus par le cahier des charges. Seul le laboratoire principal d'anatomie pathologique (cabinet SIPATH) a été équipé et relié au réseau avec les deux structures de dépistage des cancers en Auvergne. Les autres laboratoires exigent l'accord CNIL avant toute installation ;
- l'intégration directement au sein des bases de données des structures de dépistage n'a pas encore été implémentée par la société OSI-Santé, editrice du logiciel Zeus, bien que le modèle ICD et son API leur aient été soumis.

Ensuite, des développements sont encore nécessaires pour compléter objectifs techniques du réseau. En effet, bien que la mise en relation des données médicales soit effective, il manque encore le moyen de les restituer avec la plus grande qualité possible : exemptes de doublons et en rapprochant les identités des patients. Vu le caractère distribué des données, cette étape ne peut se faire de façon classique et un modèle d'identification, de comparaison et de rapprochement des identités patients adapté à l'architecture devra alors être proposé.

Chapitre 4. Gestion du patient et des données médicales pour RSCA

INTRODUCTION

L'élaboration du cahier des charges et la mise en place d'une infrastructure de grille dans le cadre du projet RSCA représentent les premières étapes de la prise en charge du patient et de ses données médicales. La réutilisation des technologies issues de globus et gLite ont permis de faciliter le déploiement de cette infrastructure, avec comme clé de voûte la technologie Gateway issue des développements de la société maat-G, partenaire du projet.

Cependant, de nombreux points sont encore à prendre en compte afin d'obtenir un système parfaitement opérationnel. Tout d'abord, il faut s'assurer que le premier objectif technique fixé par le cahier des charges, c'est-à-dire de « fournir un réseau sentinelle fortement sécurisé » soit respecté, afin de minimiser les risques de fuite d'information, ou d'utilisation frauduleuse du réseau. Les couches de sécurité des grilles fournissent les bases et outils nécessaires à assurer cette sécurité mais il est nécessaire de les ajuster au mieux à la problématique étudiée. Cette adaptation devra inévitablement être conforme au niveau de sécurité fixé par les lois informatique et libertés [8] et la directive Européenne 95/46/EC [11], les recommandations ASIP-Santé [5, 101], tout en respectant les codes de déontologie fixés par la Haute Autorité de Santé [161].

Un des points les plus importants spécifiés par les recommandations ASIP-Santé est l'utilisation des cartes de professionnel de santé (CPS) [162] dont l'étude de faisabilité a été faite en partie [3.3.1.2].

D'autres points restent sans réponse pour assurer la cohérence d'un réseau distribué et réparti de données : il s'agit principalement de l'identification du patient, déjà évoqué comme étape cruciale lors de la rédaction du cahier des charges [2.6.2]. En attendant le déploiement en France de l'identifiant national de santé (INS), dont les spécifications ont été produites en Juillet 2010 par l'ASIP [163], il faut doter le réseau d'un système d'identification des patients. Ce système devra alors s'accommoder du caractère réparti de l'architecture ainsi créée et assurer une grande fiabilité dans le rapprochement d'identités pour éviter les doubles et fausses identifications qui peuvent être extrêmement préjudiciables au patient.

En outre, un système d'identification ne suffit pas en lui-même à rapprocher de façon précise des identités distribuées. En effet, le refus, par la loi informatique et libertés et de la CNIL d'utiliser le

numéro NIR, que ce soit pour le rapprochement d'identités médicales ou bien pour son stockage dans une base de données [164] complique fortement la tâche du rapprochement d'identités.

Compte tenu de ces éléments, la seule méthode de rapprochement d'identités restant à notre disposition sera une comparaison empirique des traits d'identification d'une personne. Cette technique, issue de la fouille (ou exploration) de données est communément appelée « Data Linkage ». Elle est utilisée pour réunir deux ensembles de données sans clé commune. La qualité de cette méthode influencera alors fortement le résultat du rapprochement des identités médicales.

Une étude de cette discipline sera effectuée et une solution sera proposée en rapport avec les performances des différentes méthodes sur les jeux de données mis à disposition par le réseau sentinelle.

4.1. SECURITE DES DONNEES MEDICALES

Préambule

Au-delà de la sécurité des données médicales, la loi française (Article L1111 du code de la Santé Publique) impose dorénavant à toute structure hors du système de soin désirant stocker de l'information médicale, qu'elle soit agréée « hébergeur de données médicales ». Compte tenu du coût que représente la préparation d'un tel dossier et des délais impartis au projet, il n'a pas été envisagé, dans un premier temps, de préparer une telle habilitation.

Compte tenu de l'architecture informatique du projet, un hébergeur de données de santé n'est pas prioritaire ou limitant à l'avancement du projet. Les structures médicales n'étant pas soumises à cette réglementation, peuvent naturellement stocker les données nécessaires à leur fonctionnement.

Cette mesure a l'avantage de concentrer les problèmes de sécurité des données au niveau des structures médicales, déjà dotées de systèmes adéquats de protections physique et réseau.

C'est pourquoi la sécurité des données médicales au sein du projet RSCA se concentre sur une sécurité logicielle et réseau en empêchant ainsi l'accès à des personnes non-autorisées.

4.1.1. *Evaluation des risques*

L'évaluation des risques est une part importante et nécessaire pour la demande d'accréditation à la CNIL. Un réseau de ce type comporte des risques spécifiques qu'il est nécessaire d'étudier et de répertorier pour mieux réagir en cas d'apparition de faille de sécurité.

Comme présenté précédemment, les risques encourus sont de deux types : les risques d'ordre logiciels et les risques liés à l'utilisation du réseau avec la communication de données confidentielles sur celui-ci.

Il existe de nombreuses méthodes d'audit de sécurité disponibles : Mehari, Magerit, Cobit ou encore Ebios, cette dernière étant référence dans l'administration française. Leur comparaison dans [165] a permis de mettre en évidence la méthode Mehari [166], proposée par l'association Clusif¹ qui, par sa documentation facilement accessible et sa licence libre, a permis l'analyse des risques du projet RSCA.

4.1.1.1. *Fonctionnement de Mehari*

La [Figure 47] représente le cadre méthodologique fourni par cette méthode. Il se découpe en quatre modules : l'analyse des enjeux, des vulnérabilités, de la détection et la réduction des risques puis le pilotage de la sécurité. L'audit de sécurité consiste à répondre à environ 400 questions de type « oui/non » issues de 12 scénarios différents.

¹ Club de la sécurité des systèmes d'information Français – www.clusif.asso.fr

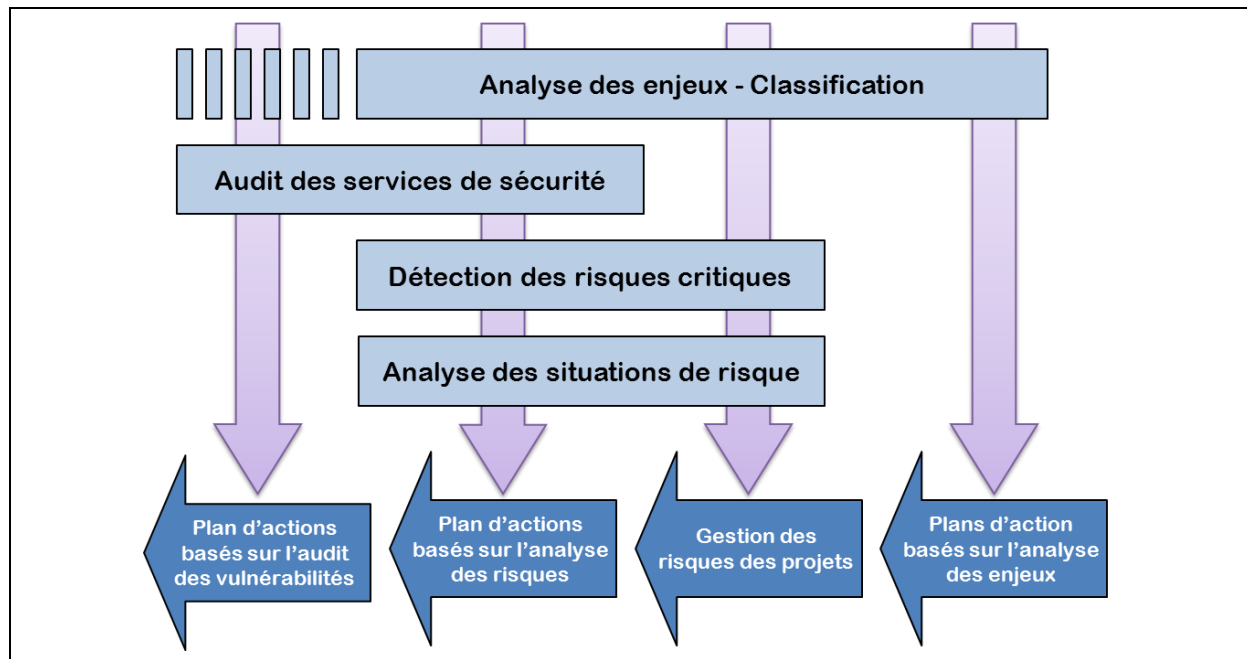


Figure 47 - Fonctionnement global de Mehari – Crédit : Clusif

4.1.1.2. Analyse des enjeux de sécurité

Cette analyse comporte, par processus métier, les risques de sécurité éventuels associés à ces processus. La [Figure 48] résume l'analyse préliminaire des risques encourus par un réseau de ce type. Ces risques sont soit :

- inhérents aux données et leur manipulation, qu'il s'agisse de l'export ou de l'import de celles-ci ;
- relatifs à l'utilisation même du réseau, de l'accès aux données comme de l'authentification des utilisateurs ;
- propres à l'administration du réseau, que ce soit du point de vue de la définition des droits des utilisateurs comme de l'identification des patients dans le réseau.

Ils sont aussi classés selon un degré (arbitraire) de gravité de leur apparition et sont déterminés par des seuils :

- Seuil 1 : Sans dommage (vert)
- Seuil 2 : Dommage important (orange)
- Seuil 3 : Dommage très grave (rouge)

En effet, certains risques sont liés aux technologies utilisées intrinsèquement (grilles informatiques) ou encore aux technologies imposées (authentification CPS) dont la marge de manœuvre en cas de défaut est beaucoup plus réduite.

Cette classification a néanmoins une importance lors de la gestion des risques : elle permet de se rendre compte des véritables vulnérabilités d'un système, en répertoriant celles qui sont internes, liées à une faiblesse des mesures de sécurité ou encore externes, dues à un défaut lié à la robustesse de la solution adoptée.

Processus	Risques	Gravité	Explication
P1: Données			
P1.1: Export de données			
R1.1.1	Interception du fichier sur le réseau		Le fichier transite sur le réseau local de la structure médicale
P1.2: Import de données			
R1.2.1	Corruption du fichier		La structure du fichier est préétabli et ne risque peu d'être modifié
P2: Utilisation			
P2.1: Authentification utilisateur par CPS			
R2.1.1	Falsification du contenu de la carte		Accès possible à des fonctionnalités non autorisées, dépend de la robustesse des cartes CPS
R2.1.2	Utilisation d'un certificat frauduleux (non autorisé)		Dépend de la robustesse de l'infrastructure PKI et des autorités de certification, a priori non critique
R2.1.3	Utilisation d'une CPS par une autre personne que son propriétaire		Nécessite l'utilisation d'un mot de passe, sous responsabilité du propriétaire
P2.2: Accès aux données médicales			
R2.2.1	Interception des données sur le réseau		La récupération de trames sur le réseau est possible mais nécessite le déchiffrement de l'information
R2.2.2	Utilisation de services sans authentification		Les services doivent avoir une politique d'accès unifiée et robuste
P2.3: Accès aux chiffres épidémiologiques			
R2.3.1	Interception des données sur le réseau		La récupération de trames sur le réseau est possible mais nécessite le déchiffrement de l'information. Données non critiques
R2.3.2	Utilisation de services sans authentification		Les services doivent avoir une politique d'accès unifiée et robuste. Données non critiques
P3: Administration			
P3.1: Définition des utilisateurs			
R3.1.1	Ajout non autorisé d'utilisateurs		Nécessite une politique fiable de sécurité et d'authentification des administrateurs
P3.2: Attribution des droits			
R3.2.1	Définition de droits non autorisé		Nécessite une politique fiable de sécurité et d'administration des droits
P3.3: Identification des patients			
R3.3.1	Interception des données d'identification		Ne permet pas d'accéder à des données sensibles (médicales)
R3.3.2	Manipulation frauduleuse du système d'identification		Nécessite une politique fiable de sécurité et d'authentification des administrateurs

Figure 48 - Récapitulatif de l'analyse des risques

Il existe cependant un point sensible au niveau de RSCA : la sécurité physique des locaux des structures médicales. En effet, la plupart des sites équipés ne sont pas à la norme recommandée par Mehari. Ceux-ci ne sont que rarement dotés d'un dispositif de sécurité adéquat, avec un réel périmètre de protection des zones sensibles.

La stratégie à adopter dans ce cas de figure n'est pas clairement établie. Il est en effet très difficile de justifier un investissement conséquent au niveau de la sécurité physique des locaux. Cependant, un moyen logiciel peut permettre de passer outre cette mesure de sécurité. En chiffrant toutes les informations stockées dans les bases de données de ces zones à risque, un vol physique des machines ne permettrait que très difficilement d'en extraire le contenu.

Deux types de chiffrements sont possibles, soit au niveau du stockage (disque dur) en encryptant toute l'information soit un chiffrement de l'information contenue dans les bases par un chiffrement/déchiffrement systématique. Tout ceci a cependant un impact négatif sur les performances globales du système qu'il s'agira alors de pondérer.

4.1.2. Éléments de sécurité requis par la CNIL

La loi Informatique et libertés [8], stipule, dans son texte, que la CNIL, autorité administrative indépendante, a pour missions principales d'informer sur cette loi et surtout de veiller à son application. Comme évoqué en [1.3.5] elle a le pouvoir d'autoriser ou non les traitements sur les données des organismes souhaitant mettre en place un tel dispositif. Dans le cadre du traitement de données nominatives à caractère médical au travers d'un réseau, les demandes d'autorisation doivent, pour être valables, respecter les contraintes de sécurité imposées par la CNIL.

La demande d'autorisation à la CNIL se compose d'une déclaration classique à laquelle s'ajoutent deux documents descriptifs :

- le premier concerne l'interconnexion des données. Il doit expliquer la nature de l'échange de données, le fondement juridique et les modalités pratiques ;
- le deuxième est entièrement dédié à la sécurité. Il a pour but de prouver que les éléments mis en place sont sans risque compte tenu du traitement sur les données mis en place.

4.1.2.1. Contenu de la déclaration

L'annexe « sécurité » de cette déclaration présente une brève description du système global, présenté ici comme un ensemble de machines s'échangeant de l'information médicale ainsi qu'un descriptif de l'application.

Au niveau général, le document comporte un ensemble descriptif de points de sécurité :

- la sécurité physique des locaux et équipements ;
- les mesures assurant la sauvegarde du système informatique ;
- la protection des équipements réseau, tant au niveau matériel (routeurs) que logiciels (pare feux).

Du point de vue de l'utilisation de l'application, divers éléments sont à renseigner :

- les bases de données et logiciels utilisés au sein de l'application ;
- les procédés techniques utilisés ;
- l'habilitation des personnes et l'authentification des utilisateurs ;
- la confidentialité des données

Par la suite, des scénarios de sécurité sont évoqués, en précisant les mesures de sécurité prises lors des différentes phases du projet : le développement de l'application, les phases de maintenance des équipements et des logiciels.

4.1.3. Utilisation de la Carte de Professionnel de Santé (CPS)

L'implémentation de l'interface pour la CPS, dont l'étude de faisabilité de l'intégration au projet a été présentée en [3.3.1.2], bien que s'inspirant de PKI, n'est pas sans difficultés au sein d'une infrastructure de grille gérée par VOMS.

Cette partie résulte d'un travail complet sur la question [167], effectué au sein du projet RSCA. Les principales difficultés rencontrées lors de l'intégration de ce système furent liées tout d'abord aux interfaces matérielles que nécessitent l'accès aux certificats stockés dans des cartes à puce. Ensuite la hiérarchie de certification fournie par le GIP-CPS [141] ne dispose pas d'une seule racine

comme il est coutume mais dispose d'une racine par classe de professionnels de santé qui, de surcroît est auto-signée. Finalement, les listes de révocation des certificats (CRL) issues des nombreuses autorités de certification (AC) ne respectent pas du tout le standard de définition proposé par PKI et sont, de ce fait, incompatibles avec VOMS en tant que telles.

4.1.3.1. Mécanismes de certification CPS

La chaîne complète de certification CPS, présentée en [Figure 49] se compose de deux grandes arborescences, une première dédiée aux tests et une deuxième à la production [141]. Au sein de ces arborescences on retrouve les mêmes autorités de certification (Anonyme ; Professionnel et Structure). La première classe désigne les établissements de santé ou une personne morale, communément appelée CPE (Carte de Professionnel d'Etablissement). La deuxième regroupe les professionnels de santé en tant que tels (médecins, pharmaciens,...). La dernière regroupe les personnes physiques œuvrant au sein d'un établissement de santé (secrétaire médicale, manipulateur, ...).

En scindant de façon claire les corps de métier de la santé, cette dichotomie, bien que facultative, a l'avantage de cloisonner au mieux les droits propres aux différentes ACs. Ainsi, vu que l'autorité qui émet les certificats « médecin » est différente de celle qui émet les certificats « établissement », il est aisé de filtrer à très bas niveau les droits d'accès de ces familles de cartes. De plus, en cas de corruption d'un des certificats « racine », ce n'est pas l'ensemble du domaine de la santé qui est alors concerné.

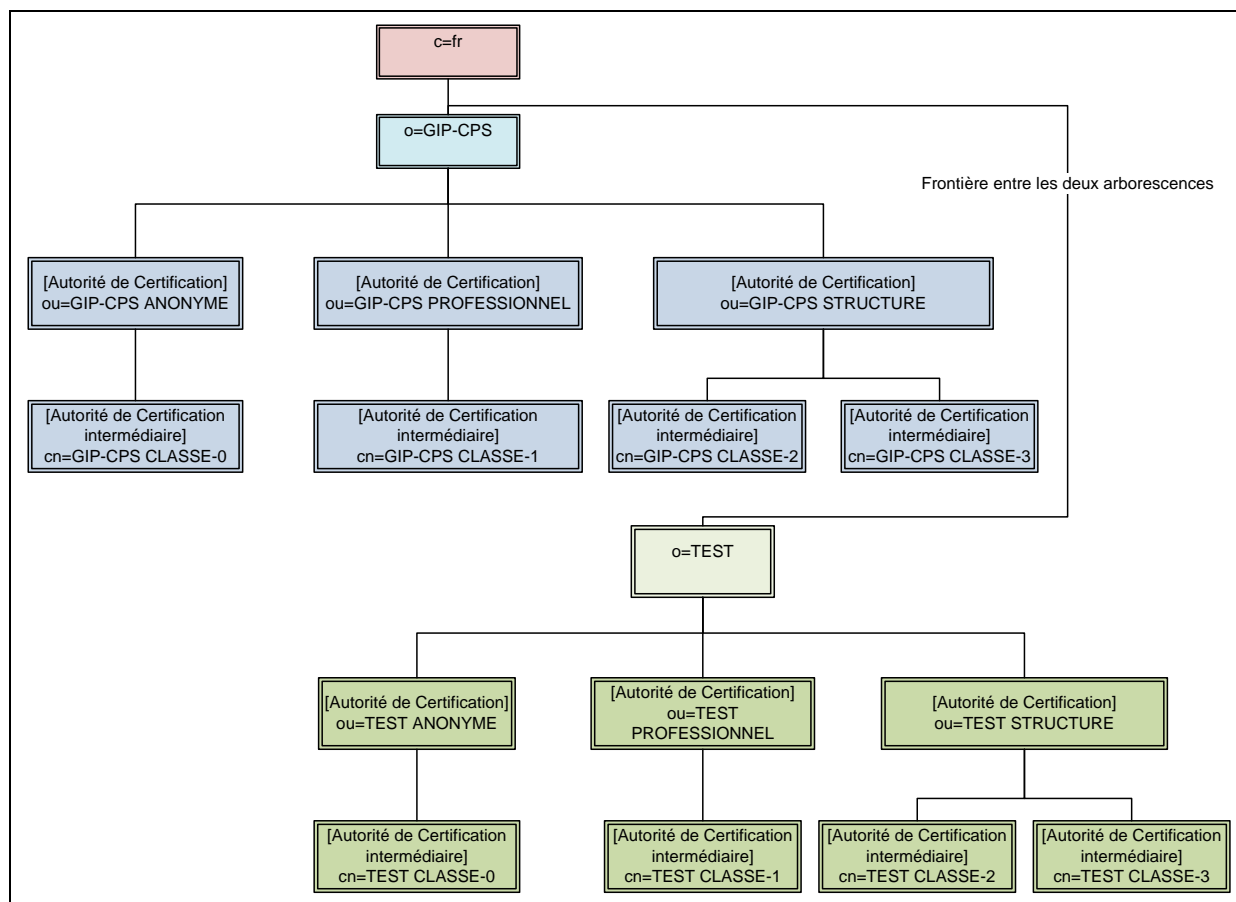


Figure 49 - Architecture des ACs du GIP-CPS

4.1.3.2. Utilisation de PKCS

PKCS (Public Key Cryptography standards) est un ensemble de spécifications proposées par les laboratoires RSA [145, 168], filiale de EMC², leader mondial du stockage, en collaboration avec les développeurs de systèmes de sécurité à travers le monde. PKCS a pour objectif de promouvoir et de faciliter le déploiement du chiffrement asymétrique par clé publique/clé privée.

PKCS a de facto repris de nombreuses technologies et standards existants pour l'élaboration des spécifications comme SSL, ou S/MIME. Les spécifications s'organisent sous la forme PKCS#XX, avec XX allant de 1 à 15 suivant le champ d'action de la norme. La [Figure 50] représente les versions les plus significatives de PKCS ainsi que leur rôle.

Dans le cadre du projet, deux d'entre elles retiennent notre attention, tout d'abord la PKCS#12, qui régit le stockage des certificats en offrant une protection par mot de passe et plus particulièrement PKCS#11 [169] qui cible les périphériques de chiffrement dont la CPS fait partie.

PKCS#11 a pour principale vocation la définition de l'API Cryptoki, qui permet d'unifier la façon d'accéder aux périphériques de sécurité, comme la CPS. Le principal avantage de cette API est son indépendance du matériel : ainsi le code créé sera valable sur l'ensemble des périphériques.

Version	Description
PKCS #1	Définition des algorithmes et du format de création des certificats. Fournit les méthodes de chiffrement/déchiffrement et vérification des signatures
PKCS #5	Chiffrement des données par mot de passe
PKCS #7	Permet de chiffrer/déchiffrer les messages dans PKI et de transmettre les certificats (en réponse à PKCS10)
PKCS #11	API fournissant une interface aux périphériques cryptographiques (Cryptoki)
PKCS #12	Permet le stockage des clés privées et clé publique (fichier .p12) en offrant une protection par mot de passe.
PKCS #15	Permet aux utilisateurs de périphériques cryptographiques de s'authentifier directement auprès des applications.

Figure 50 - Versions de PKCS

Scénario d'établissement d'une connexion sécurisée en utilisant la CPS

La [Figure 51] schématise la procédure d'établissement d'une connexion sécurisée à partir du certificat issu de la CPS en utilisant PKCS#11. Différentes étapes décrivent ce scénario :

- premièrement, le logiciel client accède au certificat X.509 stocké à l'aide de l'API Cryptoki ;
- le propriétaire entre le code de déverrouillage de ce certificat ;
- le certificat sert alors de clé pour le démarrage d'une communication chiffrée avec la Gateway ;
- la validité du certificat est vérifiée auprès de l'AC GIP-CPS correspondante à la classe de la carte CPS introduite ;
- une autorisation est alors créée sur la grille et l'utilisateur peut pleinement l'exploiter.

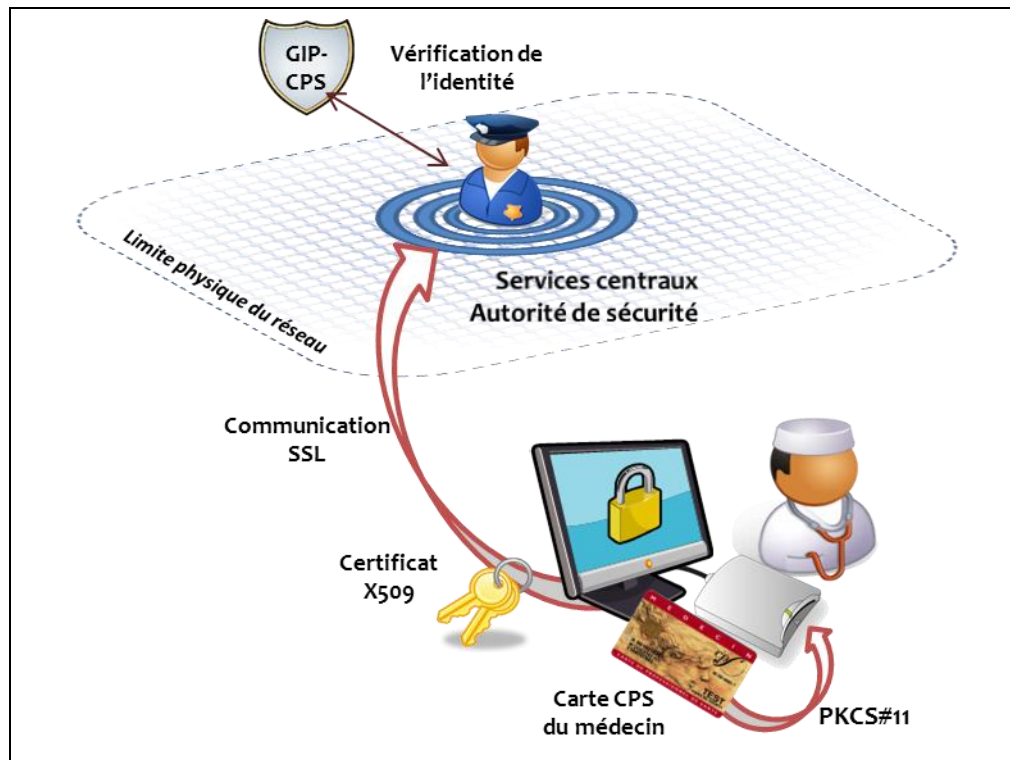


Figure 51 - Scénario d'authentification CPS

Cependant, compte tenu des spécifications CPS [143], il est nécessaire de procéder à une déconnexion de tous les services en cas « d'arrachage » de la carte de son lecteur. La procédure adoptée pour cela est de vérifier toutes les secondes la présence de la carte. En cas d'absence, toutes les ressources seront alors immédiatement détruites.

Fonctionnement de la CPS

La communication entre les applications et la carte CPS, comme le montre la [Figure 52], suit un modèle en couches où figure naturellement PKCS#11. Ainsi, une succession de pilotes matériels puis CPS (fournis par le GIP-CPS) se greffent au lecteur de cartes. Par-dessus se greffe PKCS, qui unifie et facilite l'accès au matériel depuis l'application.

Ainsi, pour le projet, seule la couche supérieure est à implémenter. Le pilote de la CPS étant mis à disposition pour la plupart des systèmes d'exploitation (macOSX, Linux, Windows).

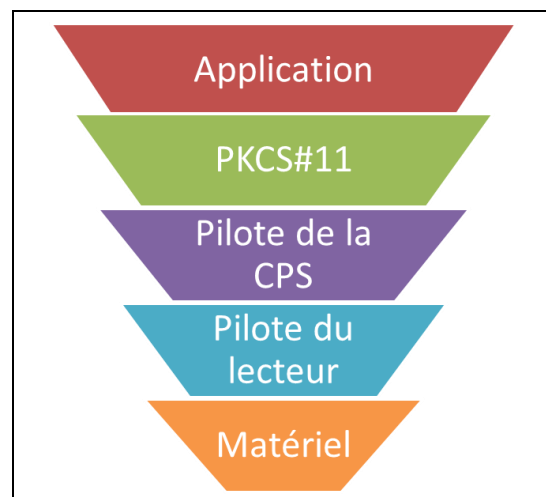


Figure 52 - Schéma en couches de la communication avec la CPS

Cependant, l'utilisation brute de PKCS#11 n'est pas aisée en tant que tel et les applications de base fournies par le GIP-CPS ne permettent pas en l'état d'être utilisables par une autre application. Dans le cas du projet RSCA, il est nécessaire d'accéder au certificat contenu dans la carte, le déchiffrer puis le transmettre par un canal sécurisé au service d'authentification sur la grille (VOMS).

Utilisation de Network Security Services (NSS)

Pour cela, NSS [170] est un ensemble de bibliothèques proposé par Mozilla pour le développement de la sécurité au niveau des architectures Client/Serveur. NSS est notamment utilisé par des bibliothèques comme PKCS#11, SSL et des logiciels (Apache, Firefox). Il a notamment été repris par Java/J2EE sous la forme du paquetage `java.security`.

Ainsi, l'utilisation conjointe de NSS et de Java permet d'implémenter le nécessaire au développement d'une solution client accédant et décodant le certificat.

4.1.3.3. Utilisation des ACs GIP-CPS avec VOMS

La dernière étape, côté serveur, est d'installer la chaîne de certification issue du GIP-CPS au sein de VOMS. En effet, VOMS a besoin, pour garantir la validité du certificat, d'être en présence de l'intégralité de la chaîne de certification afin de vérifier l'authenticité d'un certificat.

Scénario d'authentification

L'application d'authentification, présentée précédemment, établit la connexion sécurisée puis transmet le certificat au serveur VOMS distant. Celui-ci, va alors évaluer la validité du certificat par rapport aux certificats racine de l'AC émettrice.

Ensuite, suivant les droits internes à VOMS, il va autoriser, ou non l'établissement du proxy grille qui débloquera l'accès aux services. Il ne s'agit pas cependant de donner l'accès à RSCA à tout possesseur d'une carte CPS valide conformément au cahier des charges.

Intégration des listes de révocation de certificats (CRL) dans VOMS

La partie la plus délicate ici est d'intégrer de façon automatique les CRLs par VOMS. Le GIP-CPS, via son annuaire [144], publie de façon régulière les CRLs associées aux certificats CPS émis par les différentes ACs. Le format standard de publication d'une révocation d'un certificat, pour éviter qu'une tierce partie puisse elle-même émettre une révocation de façon frauduleuse, est de signer la révocation par la clé privée de l'AC émettrice. Ainsi, la vérification de la signature permet d'authentifier la demande de révocation.

Cependant, ce mécanisme n'est pas respecté au sein de la chaîne de certification du GIP-CPS. Ainsi, les CRLs ne sont pas signées par le certificat de l'AC concernée mais par celle immédiatement supérieure au sein de la hiérarchie. Cette méthode n'est pas du tout standard et n'est pas compatible avec VOMS sans modification du code source du logiciel. Pour des raisons pratiques, la vérification des CRLs a été désactivée.

Puisque l'autorisation est aussi effectuée manuellement par les administrateurs dans VOMS, l'absence de cette fonctionnalité ne remet pas en cause la sécurité du réseau.

4.2. L'IDENTIFICATION DU PATIENT

Bien que ce problème ait été résolu dans de nombreux pays (Belgique/Espagne) avec la création d'identifiants disposant d'un cadre légal moins restrictif, l'identification du patient (ou des personnes), en France n'est pas encore aboutie. En effet, bien que disposant depuis 1946 [171], avec la création de l'INSEE, du numéro de sécurité sociale, la loi informatique et libertés impose que cet identifiant ne soit pas utilisé pour le croisement de fichiers, même au sein de structures médicales [164]. Bien entendu, des dérogations sont accordées pour certains organismes, liés à l'emploi, aux prestations sociales ou médicales, mais ne sont en aucun cas généralisables.

4.2.1. *Les enjeux de l'identification*

Au sein des différents pays, d'après [172], il est possible de distinguer 3 cas de figure sur les systèmes d'identification des patients :

- les pays où l'identification est gérée de façon nationale, largement établie et permet l'interopérabilité nécessaire à la création du dossier médical partagé. En font partie le Danemark, Pays-Bas, Belgique ou le Royaume-Uni ;
- les pays où le système est géré à plus petite échelle, de la région ou de la communauté (Espagne et Italie) ou au sein d'états dans les pays fédéraux (Canada, Australie) mais qui disposent d'un système d'identification globalement efficace ;
- les pays indécis, comme les Etats-Unis ou la France, qui disposent depuis longtemps d'un identifiant national (SSN pour les Etats-Unis), (INE pour la France) mais ne peuvent l'utiliser que partiellement pour le système de soin. Aux USA, les risques de fraude au numéro de sécurité sociale sont très nombreux puisque, contrairement à la France, il n'existe pas de carte à puce de santé similaire à la carte Vitale. Une proposition de loi, HIPAA¹ [173], présentée en 1996 vise à l'établissement de standards pour les transactions médicales et la définition d'un nouvel identifiant, le NPI².

La France, quant à elle, pourrait entrer dans la première catégorie, mais le refus par la CNIL de la généralisation de ce numéro a compliqué fortement la problématique. L'ASIP-Santé a justement présenté l'identification du patient comme un outil stratégique.

4.2.1.1. *Risques liés à l'identification*

Le rôle central de l'identification du patient dans le système de soin est encore renforcé lorsque l'on évoque les risques de la mauvaise identification. On peut distinguer 3 grands types de risques [174] :

- la double identification, i.e. deux identifiants pour la même personne ;
- la fausse identification, i.e. deux personnes avec le même identifiant ;
- la non-identification lorsque l'attribution d'un identifiant n'est pas possible (prérequis non satisfaits).

¹ Health Insurance Portability and Accountability Act

² National Provider Identifier

L'objectif d'un système d'identification est justement de minimiser ces trois risques. La fausse identification étant la plus dangereuse de toutes, car elle amènerait à une situation délicate vis-à-vis des deux patients confondus, surtout en cas de pathologie grave.

4.2.1.2. Identifiant National de Santé

Pour tenter de résoudre le problème de l'identification des patients en France, dans le cadre plus large qu'un seul système de soin, l'ASIP-Santé a proposé une spécification de deux nouveaux identifiants, baptisés INS-C et INS-A [163] en collaboration avec l'ANSSI [175].

L'INS-C, pour « Calculé » est dérivé du numéro de sécurité social de façon déterministe. L'INS-A est un simple numéro à 12 chiffres tiré de façon aléatoire.

Pratiquement, [176] l'INS-C est calculé à partir des informations contenues dans la carte vitale du patient : les prénoms, date de naissance et le numéro de sécurité social.

Exemple de calcul de l'INS-C :

Jean Dupond, né le 01/01/1950 avec comme INE : 150016321200154

La première étape consiste à créer une chaîne de 20 caractères contenant 10 caractères pour le prénom, 6 pour la date de naissance(AAMMJJ) et les 13 derniers pour l'INE. Ici, notre patient aura comme chaîne : **[JEAN 500101150016321200154]**

A noter qu'en cas de caractères manquants, ceux-ci sont remplacés par des espaces. Ensuite le tout est « haché » en utilisant l'algorithme SHA-256. L'utilisation de cette fonction de hachage empêche de remonter aux informations qui ont été utilisées pour créer le condensat¹.

Résultat du hash : b724fe96515eb00d52db3597592ebdc7f83af52b0bb9f9dc8472089d086394

Ensuite le nombre est converti en numérique (chiffres de 0 à 9). Seuls les 64bits de poids fort sont pris en compte (**b724fe96515eb00d**) :

Résultat de la conversion : 13 196 952 729 666 105 357

Une fois cette « empreinte » calculée, il faut créer un numéro de contrôle de la somme (checksum) qui est le résultat de la division euclidienne de l'empreinte par 97 (nombre premier) que l'on ajoute au numéro :

Résultat final : 13 196 952 729 666 105 357 62

L'INS-C comporte de nombreux inconvénients : il est prévisible, long et avec un taux de collision non nul entre deux nombres. De plus, sa saisie n'est pas évidente (22 caractères). Ainsi, en attendant la généralisation de l'attribution de l'INS-A à l'ensemble de la population, ce numéro pourra servir de manière transitoire.

INS-A

L'INS-A, pour INS-Anonymisé, futur identifiant définitif de santé, proposé normalement début 2011 permettra de combler les inconvénients de l'INS-C. La conception de cet identifiant vérifie ces attributs :

¹ Résultat d'une fonction de *hash* comme sha-256

- unicité : un seul INS pour chaque personne tout au long de sa vie ;
- non-signifiante : la connaissance de l'INS ne permet pas de déduire des informations sur la personne ;
- sans doublon ni collision : un et un seul INS par personne ;
- non prédictibilité : la connaissance du NIR ou des traits d'identité de la personne ne permet pas de déduire l'INS;
- indépendance : l'INS-A ne permet pas de remonter au NIR de la personne.

Seulement, l'INS-A ne sera pas généralisé immédiatement à l'ensemble de la population. L'attribution de l'identifiant se faisant au cas par cas, à l'ouverture de leur dossier médical partagé. Le système de soin devra ainsi supporter une phase transitoire composée de l'INS-C et de l'INS-A simultanément.

Utilisation de l'INS-C et de l'INS-A

Bien qu'il soit difficile de remonter aux traits d'identification d'une personne en partant de l'INS-C, il présente toujours un petit risque puisque cet identifiant est calculable.

Dans le cas de RSCA, l'INS pourrait servir de base à l'identification. Cependant, aucun cadre légal, du point de vue de la CNIL n'encadre l'utilisation de ce numéro. Il est impossible de savoir, à l'heure actuelle, si ces numéros pourront être utilisés pour un réseau de ce type et permettront de rapprocher des identités distribuées. De plus, bien que l'INS-C soit facilement calculable, la CNIL n'a toujours pas statué sur sa possible utilisation dans un réseau d'échange de données comme elle l'avait pourtant fait avec le NIR.

Afin de respecter les délais impartis au projet, il n'est pas possible d'attendre la mise en place définitive du cadre technique et du statut légal de ces éléments. Il est cependant indispensable de connaître leur fonctionnement et permettre leur prise en compte ultérieurement.

4.2.2. Contraintes d'un système d'identification

Compte tenu des éléments précédents, la création d'un autre système d'identification est nécessaire. Il convient alors d'énumérer les différentes contraintes que comporte alors un système de ce type.

Celles-ci ne sont forcément pas très éloignées des contraintes de création de l'INS. Nombre d'entre elles sont directement liées à la loi Informatique et Libertés. A savoir :

- La non-signifiante :

Il doit être impossible, à partir d'un identifiant, de retrouver quelque information sur la personne identifiée. Il est soit possible d'utiliser des méthodes d'anonymisation (irréversible) ou de pseudonymisation (irréversible) d'identifiants mais la solution la plus aisée est d'utiliser un nombre pseudo-aléatoire.

- L'unicité :

Un identifiant doit être unique, sans possibilité de doublon ou de collision d'identifiant.

- L'évolutivité et la souplesse :

Le système doit permettre une flexibilité qui est propre à l'identité évolutive d'un patient dans le système de soin. Le caractère distribué du réseau doit se doter d'un système d'identification adapté à cette architecture particulière.

- La sécurité :

L'identifiant ne doit pas être en tant que tel une clé d'accès au système. Une usurpation d'identifiant ne doit pas permettre de retrouver l'intégralité des données concernant le patient.

- L'isolation :

La corruption d'un élément du réseau ne doit pas permettre l'accès à l'ensemble des données de celui-ci.

- L'indépendance :

Le système d'identification doit être indépendant des données gérées par le réseau. Pour des raisons aussi de sécurité, l'entité qui gère l'identification ne doit pas avoir en sa possession les données traitées. Cette contrainte permet, en outre d'externaliser ce système du monde médical.

4.3. MODELE D'IDENTIFICATION DYNAMIQUE ET DISTRIBUE DU PATIENT

L'analyse préalable de la situation de l'identification du patient en France a permis de comprendre en quoi il n'est pas possible de réutiliser l'existant pour créer un système de gestion de données médicales partagées en utilisant les technologies de grille. Vu l'aspect prépondérant de cette partie du réseau, il est nécessaire de procéder au développement d'une nouvelle solution qui puisse à la fois convenir à l'architecture adoptée mais surtout de se conformer aux contraintes et dispositions légales évoquées précédemment.

4.3.1. Présentation de la solution adoptée

La solution proposée consiste alors à créer, autour d'une identité virtuelle du patient, une fédération d'identifiants issus des structures partenaires du réseau.

Ces identités virtuelles seront gérées par un service externe à toute structure médicale qui ne se chargera uniquement du lien entre les identifiants. La question de l'hébergement de données médicales ne se pose alors pas.

4.3.1.1. Eléments de sécurité

La sécurité sera assurée à plusieurs niveaux :

- par l'utilisation systématique du chiffrement des identifiants. Seul le site concerné doit être capable d'exploiter/déchiffrer un identifiant ;
- par le chiffrement de la transmission de l'identifiant en utilisant les mécanismes de clé publique/clé privée.

Ce dernier point a l'avantage, à partir d'un identifiant non chiffré, de créer des doubles chiffrés qui seront stockés dans les bases des systèmes ayant chiffré l'identifiant. Par ce biais, la corruption d'un site ne permet pas d'avoir accès à l'ensemble du réseau.

La [Figure 53] schématise la relation entre l'identifiant nominal d'un patient et ses différentes versions chiffrées stockées dans les bases de données distantes.

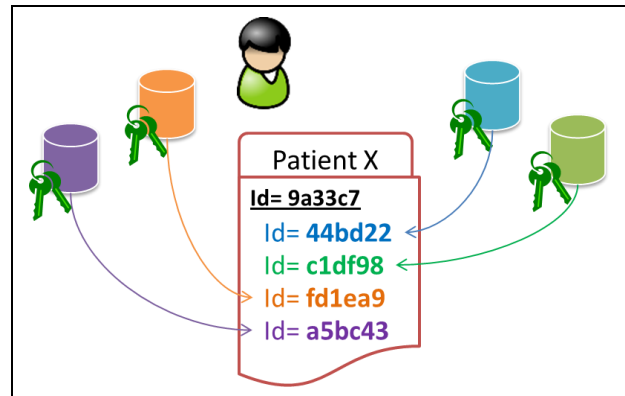


Figure 53 - Identifiants et versions chiffrées

Pour un patient donné et à partir d'un identifiant A qui lui est attribué, chaque site $S_1..S_N$ disposant d'une clé privée $S_{ka1}..S_{kaN}$ et d'une clé publique $S_{kb1}..S_{kbN}$, peut stocker une version chiffrée de l'identifiant A en utilisant la fonction de chiffrement $C : C(A, S_{kb1})..C(A, S_{kbN})$ à l'aide de sa clé publique. De cette manière, seule leur clé privée pourra déchiffrer cet identifiant, en utilisant toujours la même fonction de chiffrement C .

Pour la transmission d'un identifiant chiffré depuis un site i vers un site j , deux étapes sont alors nécessaires :

- déchiffrer l'identifiant stocké $C(A, S_{kbi})$ avec sa clé privée S_{kai} : $C(C(A, S_{kbi}), S_{kai})$
- le re-chiffrer avec la clé publique du destinataire S_{kbj} : $C(C(C(A, S_{kbi}), S_{kai}), S_{kbj})$

La transmission est alors possible, seul le destinataire avec sa clé privée S_{kaj} pourra déchiffrer l'identifiant.

4.3.1.2. Fédération d'identifiants

L'identité du patient au sein des structures médicales est, hors INE, le plus souvent isolée, sans possibilité de rapprochement immédiat en utilisant une clé commune.

La [Figure 54] représente clairement ce problème : un patient dispose le plus souvent d'un identifiant interne à la structure mais celui-ci diffère d'un établissement à l'autre, sans possibilité de les relier.

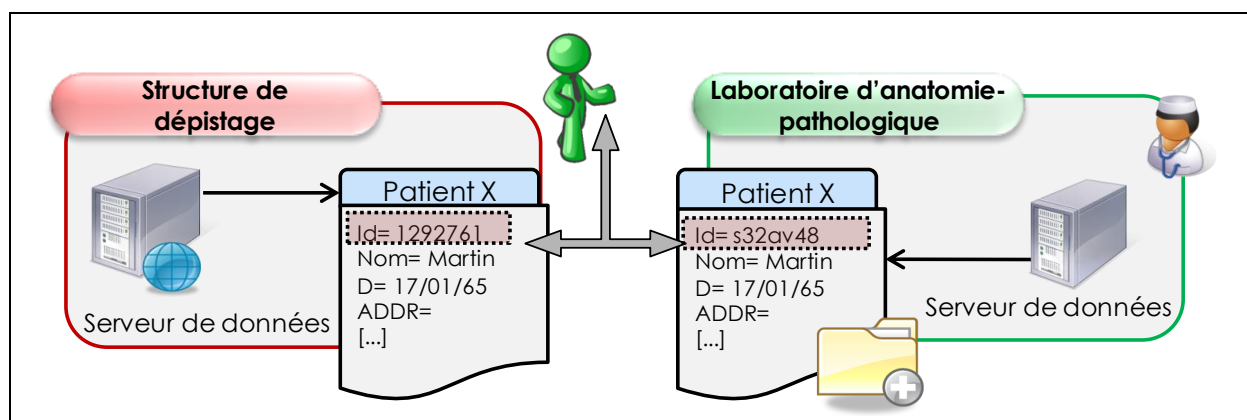


Figure 54 - Problème d'identification

L'objectif ici est alors de regrouper, en plus des identifiants globaux, ceux issus des structures incluses dans le réseau. Cette fédération d'identifiants permettra à une structure centrale d'effectuer les rapprochements nécessaires à la réalisation des objectifs du réseau.

Exemple, pour le patient X de la [Figure 54] : autour d'un identifiant global du réseau sera ajouté un couple (N° de structure, N° de dossier) qui permettra à cette identité d'être reliée aux différentes structures qui possèdent des données le concernant : voir [Figure 55].

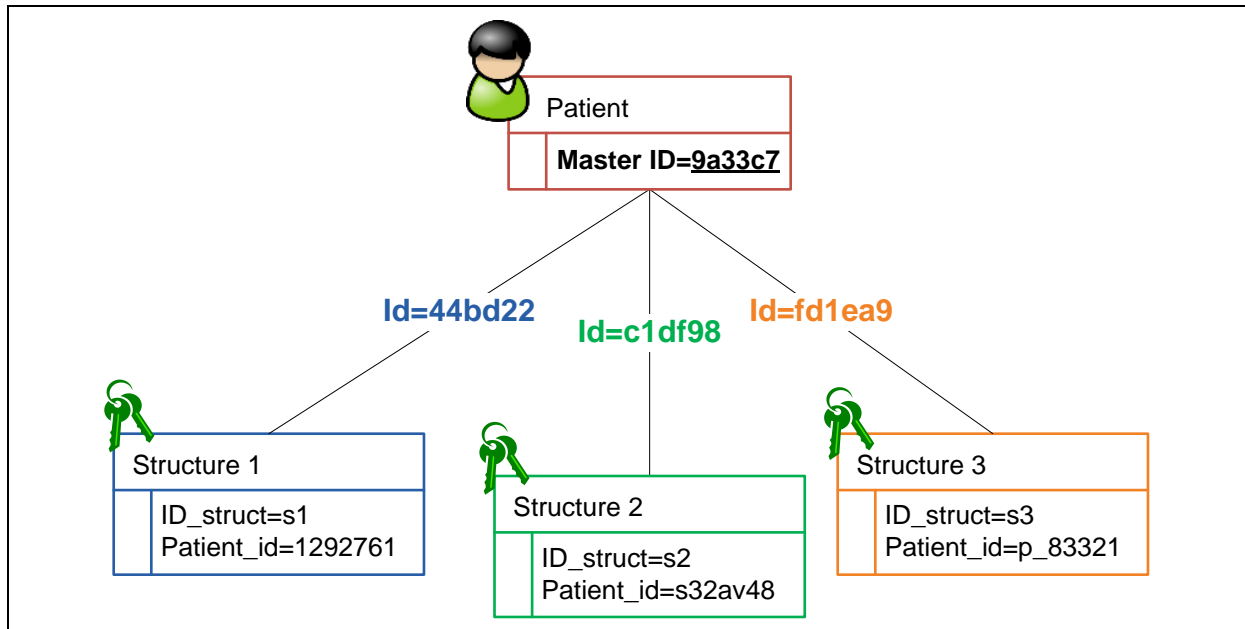


Figure 55 - Arbre d'identification pour un patient

4.3.2. Description des scénarios d'identification

4.3.2.1. Choix de l'identifiant

La solution proposée doit s'appuyer sur un identifiant complètement anonyme et généré aléatoirement, sans aucun rapport avec le patient. Une des manières les plus simples de générer ce type de numéro est d'utiliser le système uuid¹, proposé par P. Leach [177] et présenté sous la norme RFC4122. Ce système génère un nombre aléatoire sur 128bits avec 122bits aléatoires, soit $2^{122} \approx 10^{36}$ possibilités. Les probabilités de collision sont extrêmement faibles et l'entropie de l'algorithme en fait un identifiant statistiquement unique.

L'implémentation de l'algorithme au sein des plateformes informatiques est très répandue, que ce soit dans les systèmes, langages de programmation ou même bases de données. Il est disponible nativement sous tout système unix/linux/osx via la commande **uuidgen**, implémenté dans java sous le paquetage **java.util.UUID** ou intégré de façon native dans la base de données mysql par la fonction **UUID()**.

¹ Universal Unique Identifier

4.3.2.2. Scénarios d'identification

Ces différents scénarios sont au nombre de cinq :

1. Ajout des données au réseau par un fournisseur
2. Requête de type nominative
3. Requête de type statistique (anonyme)
4. Comportement du service central d'identification
5. Réponse aux requêtes par le service central

1ère phase : Ajout de données dans le réseau sentinelle par un fournisseur de données	
Synopsis	Un utilisateur souhaite intégrer des données au réseau sentinelle
	<ul style="list-style-type: none"> • Il s'identifie sur le serveur grille de son établissement relié au réseau sentinelle • Il « pousse » ses données sur ce serveur
Le serveur grille reçoit les données : Pour chaque entrée reçue, il vérifie auprès du serveur d'identification central l'existence du patient concerné : 3 cas sont à distinguer :	
Cas 1	<ul style="list-style-type: none"> • Le patient n'existe pas : taux de rapprochement < limite_bas : • Il insère ce patient dans la base de données en créant un nouvel identifiant
Cas 2	<ul style="list-style-type: none"> • Si ce patient existe déjà (i.e. il est possible de rapprocher l'identité à un autre patient avec un taux de confiance suffisant) -> taux > limite_haut • Alors il récupère l'identifiant de ce patient sur le réseau et l'attribue au patient dans la base
Cas 3	<ul style="list-style-type: none"> • Sinon (le taux de confiance est insuffisant) limite_bas < taux < limite_haut • Le serveur place le patient en attente d'une validation manuelle d'un opérateur en comparant visuellement les fiches patient

Cette première phase fait appel à une notion de taux de rapprochement et de seuils limites qui seront discutés au cours de la partie suivante [4.4].

L'insertion de l'identifiant du patient dans le réseau sentinelle se fait en chiffrant les données à l'aide de sa clé publique. Ainsi il n'est pas possible de récupérer l'identifiant « clair » sans posséder la clé privée du serveur concerné.

2ème phase : Requête des données sur le réseau par les associations (type nominatif)	
Synopsis	Une association souhaite récupérer les données sur un patient
	<ul style="list-style-type: none"> • L'utilisateur s'identifie sur le serveur grille de son établissement relié au réseau sentinelle • Il recherche le patient dans son logiciel métier local et effectue une mise en correspondance avec celui stocké dans la base de son serveur grille • Si le patient n'existe pas il faut insérer le patient en procédant de la même façon qu'en phase 1
Une fois l'identifiant récupéré depuis le serveur de grille:	
	<ul style="list-style-type: none"> • Il est maintenant déchiffré via la clé privée de l'association • Il est de nouveau chiffré avec la clé publique du serveur central d'identification • La requête est maintenant envoyée au serveur d'identification (voir phase 4)
Une fois la requête transmise, à chaque nouveau message :	

	<ul style="list-style-type: none"> • Il récupère les résultats préalablement chiffrés par sa clé publique • Il déchiffre le contenu du message au moyen de sa clé privée • Il insère les données dans la base de données en prenant soin de chiffrer à nouveau l'identifiant au moyen de sa clé publique.
Une fois les données insérées, le logiciel métier de l'association peut alors requêter son serveur de grille local pour récupérer les informations.	
Passé un délai ou une fois les informations récupérées par le logiciel métier, toutes les données (mises à part l'identifiant) sont supprimées du serveur de grille car elles sont en doublon (par respect des contraintes des fournisseurs)	

3ème phase : Requête des données sur le réseau par la santé publique	
Synopsis	Un utilisateur souhaite récupérer des données statistiques
	<ul style="list-style-type: none"> • Il s'identifie sur le serveur grille de son établissement relié au réseau sentinelle • Il transmet une requête au serveur central avec ses paramètres (voir phase 4) • Il attend la réponse et récupère les résultats une fois disponibles

4ème phase : Serveur d'identification	
Synopsis	Le serveur d'identification central reçoit une demande de données
	<ul style="list-style-type: none"> • Il reçoit le message • Il déchiffre son contenu grâce à sa clé privée
Suivant le contenu du message deux cas sont à distinguer : Soit c'est une requête statistique :	
	<ul style="list-style-type: none"> • Il requête alors les différents serveurs de données concernés auxquels a accès l'émetteur de la requête • Pendant un temps défini il attend les résultats • A la réception d'un message, il enregistre les résultats dans une liste • Une fois que les serveurs ont tous répondu ou que le temps a expiré, le message est envoyé à l'expéditeur préalablement chiffré avec sa clé privée
Soit c'est une requête de données nominatives	
	<ul style="list-style-type: none"> • Il récupère l'identifiant du patient (préalablement déchiffré) • Pour chaque serveur de données « X » disponible et concerné par la requête : <ul style="list-style-type: none"> ○ Il chiffre l'identifiant avec la clé publique du serveur « X » ○ Il transmet la demande à ce serveur (phase5) ○ Pendant un temps défini, le serveur attend les résultats des différents serveurs (voir phase 4) • A la réception d'une réponse : <ul style="list-style-type: none"> ○ Il déchiffre le message avec sa clé privée ○ Il transmet ce message à l'expéditeur en prenant soin de le re-chiffrer avec la clé publique de l'expéditeur

5ème phase : Réponse aux requêtes par un fournisseur	
Synopsis	Le serveur de données reçoit une requête
	<ul style="list-style-type: none"> • Il reçoit le message • Il déchiffre son contenu grâce à sa clé privée • Il vérifie les droits d'accès de l'émetteur de la requête
Suivant le contenu du message deux cas sont à distinguer : Soit c'est une requête statistique :	
	<ul style="list-style-type: none"> • Il exécute la requête et renvoie un résultat chiffré au serveur central
Soit c'est une requête de données nominatives	
	<ul style="list-style-type: none"> • Il récupère l'identifiant chiffré et le compare à sa base de données locale • Il renvoie les données demandées si le patient existe en prenant soin de chiffrer la réponse avec la clé publique du serveur central

Avantages et inconvénients de la méthode

Cet ensemble de scénarios a le principal avantage d'amener un haut niveau de sécurité : les données ne sont jamais transférées sans chiffrement. De plus, jamais l'identifiant du patient n'est « en clair » dans le réseau sentinelle. Cette mesure assure en outre que l'identifiant ne peut être mis en relation sans avoir accès à une clé privée d'un site du réseau.

Le rôle central du serveur d'authentification permet d'en faire un nœud centralisé de contrôle de l'accès et le traçage des utilisations, ce qui est une recommandation de l'ASIP-Santé.

Cette agrégation d'identifiant a aussi la possibilité d'intégrer d'autres identifiants ou sources d'identification, comme l'INS-A une fois qu'il sera disponible. De plus, en cas de double ou fausse identification, une modification de la liste des identifiants associés au patient permet de résoudre le problème.

Enfin, le dernier avantage est de créer un réseau où toutes les données sont clairement identifiées, ce n'est pas un simple dépôt non structuré de l'information.

Cependant, cette méthode a des inconvénients non négligeables : le système par chiffrement/déchiffrement peut devenir conséquent lors de la montée en puissance du réseau sentinelle. De plus, la première phase nécessite, par patient, une comparaison systématique avec l'ensemble des patients déjà existants. Des méthodes d'optimisation seront nécessaires.

Enfin, la première phase de rapprochement nécessite une véritable recherche sur la méthode de comparaison de patient, à savoir qu'est ce qui définit une relation entre deux dossiers du même patient ? Comment mesurer cette similarité ? Et aussi quelles sont les valeurs des seuils à fixer **limite_bas** et **limite_haut** ?

La [Figure 56] résume les différentes étapes du système d'identification, à partir de plusieurs sources des données concernant un même patient. Les identifiants colorés indiquent les doubles chiffrés de l'identifiant nominal du patient qui ne sont pas visibles « en clair » sans déchiffrement au préalable.

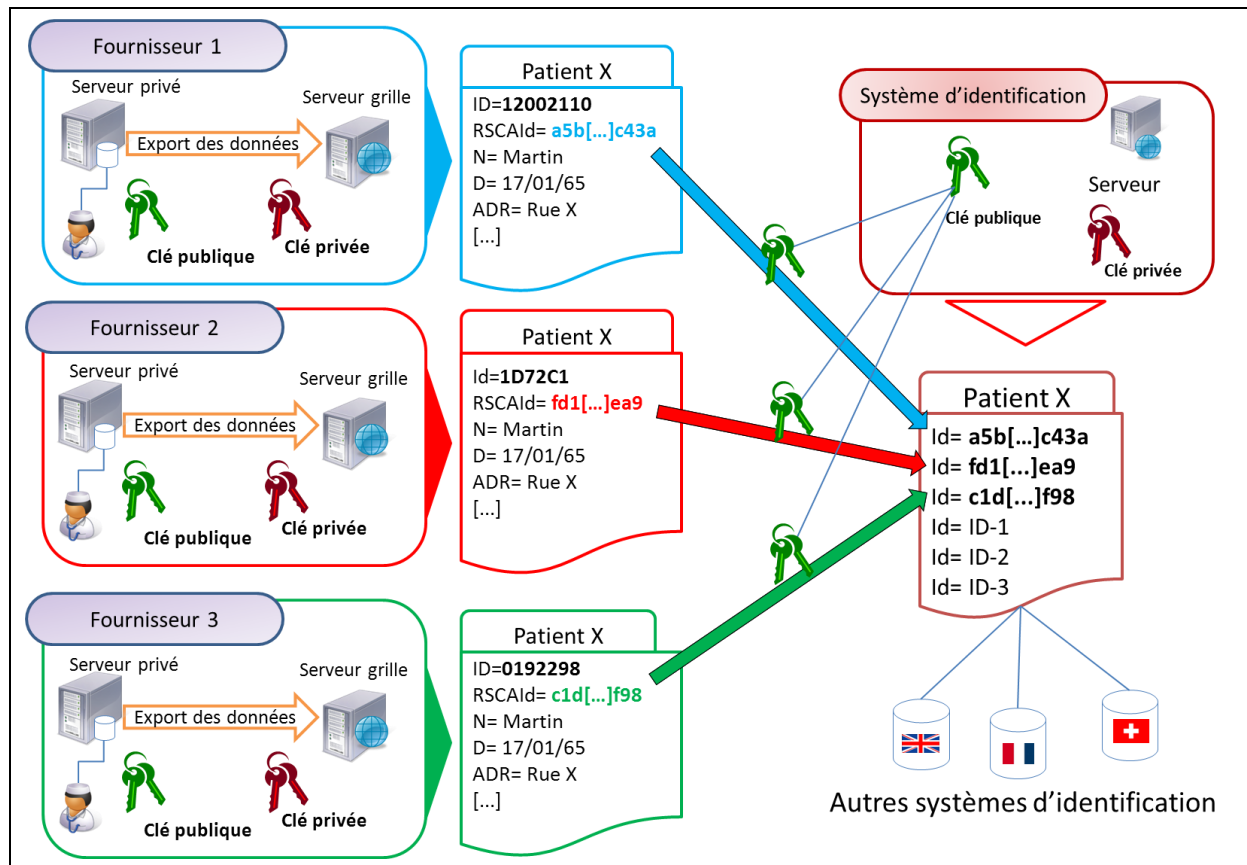


Figure 56 - Fonctionnement du système d'identification

Cette méthode a aussi l'avantage de séparer en deux couches bien distinctes l'identification des patients et l'utilisation du réseau.

De cette manière la sécurité des informations est un peu plus protégée : une requête en identification ne comporte aucune donnée médicale et une requête en données médicales n'a pas besoin des traits d'identification du patient mais seulement de l'identifiant anonyme et chiffré du réseau.

4.4. RAPPROCHEMENT D'IDENTITES MEDICALES DISTRIBUEES

Comme évoqué dans la précédente partie, la création d'une « identité » du patient au sein du réseau sentinelle est loin d'être une chose triviale. Une fois le modèle d'identification du patient créé, une autre grande partie des développements nécessaires à la mise en œuvre du réseau sentinelle concerne le rapprochement des patients au travers de l'infrastructure créée.

D'un point de vue théorique, l'interconnexion des bases de données, trivialement appelée « croisement de fichiers » ne peut se faire qu'en deux étapes en l'absence d'un identifiant commun :

- premièrement, en créant l'infrastructure qui permet « physiquement » l'accès distant, avec tous les garde-fous nécessaires pour assurer la sécurité ;
- deuxièmement, en reliant les informations communes contenues dans ces bases. Et cette étape n'est possible, pour les raisons expliquées précédemment [4.2] qu'en comparant de façon empirique les données.

4.4.1. *Enjeu du rapprochement des patients pour l'épidémiologie*

L'enjeu de cette étape est central pour le réseau sentinelle. L'interconnexion simple des bases de données médicales ne permet évidemment pas d'exploiter de façon fiable les données sans passer par un long travail de nettoyage et de dédoublonnage de l'information.

En effet, pour qu'une enquête épidémiologique ait des résultats fiables il faut absolument que les données en entrée soient les plus propres possible. Le nettoyage des jeux de données est une étape préalable fastidieuse et souvent plus coûteuse que l'analyse en elle-même des données. C'est pourquoi les analyses épidémiologiques sur le cancer ne sont souvent disponibles qu'après 3 années de travail.

Une étude menée par Friedman [178] a montré jusqu'à 27% d'erreurs d'identification de patients au sein de trois bases de données hospitalières de 100.000 patients sur le même site.

Le but du rapprochement des identités médicales distribuées est alors de réduire drastiquement le temps de disponibilité des données pour l'épidémiologie. L'objectif final étant de les mettre à disposition en temps quasi réel, c'est-à-dire peu de temps après l'intégration des données au réseau.

4.4.2. *Les techniques de rapprochement de données*

4.4.2.1. *Origine du rapprochement de données*

La technique de rapprochement de données issues des bases distinctes sans identifiant commun est une discipline à part entière appelée « data linkage » ou « record linkage ». A cela le préfixe « medical » peut être ajouté lorsque le rapprochement porte sur des données de santé. On peut alors parler de « medical data linkage » ou plus exactement « medical record linkage ». La première mention du « record linkage » a d'ailleurs été évoquée par H.L. Dunn en 1946 [179] dans un contexte de santé publique. Il présente le « record linkage » par analogie au livre de la vie, où chaque événement constitue un enregistrement ou une page de ce livre. Le « record linkage » serait alors la mise en relation de ces pages pour former le livre.

Par la suite, Newcombe [180] en 1967, puis Acheson, en 1969 [181] ont effectué des travaux pionniers dans le contexte médical. Cela a amené les premiers systèmes de « medical record linkage » au Canada puis au Royaume-Uni dès les années 1960.

Les premières applications ont été de rapprocher une identité dans des bases hospitalières qui ne disposaient pas d'identifiant commun, par la suite le dédoublonnage de jeux de données a bénéficié de ces techniques en cherchant des enregistrements similaires dans des bases de données.

4.4.2.2. *Erreurs fréquentes de rapprochement d'identités*

Si l'on considère deux ensembles de données R et S comportant une clé (ou identifiant) commune, une simple jointure naturelle (notée \bowtie) du type $R \bowtie S$ suffirait à croiser les fichiers. Cependant, dans le cas où un identifiant serait mal saisi par un opérateur, il se pourrait que cette jointure ne vérifie pas *explicitement* l'identité de la personne considérée.

Dans le cas où, cette clé commune n'existerait pas, mais qu'un attribut $R(name)$ et $S(nom_de_famille)$ soit commun entre les deux enregistrements il est alors possible d'effectuer

une equi-jointure de type $R \bowtie S_{name=nom_de_famille}$. Cependant, cette méthode est très dangereuse car elle va relier toutes les personnes ayant le même nom entre les deux ensembles.

On comprend alors qu'une solution en algèbre relationnelle ne suffit pas en elle-même. Il est alors nécessaire de procéder à la comparaison de *l'ensemble* des attributs communs simultanément entre deux enregistrements.

Par-dessus se greffe un grand nombre d'erreurs probables sur les enregistrements à comparer. On distingue plusieurs types d'erreurs :

- intrinsèques aux enregistrements « ressemblants » : homonymes, jumeaux ou résidents à la même adresse ;
- liées à la conception même des systèmes d'information : erreur de codification des caractères : « De Vlieger \neq De_Vlieger », limitation des champs à un certain nombre de caractères « Charpentier \neq Charpent »
- humaines (le plus fréquemment) d'où on distingue trois cas de figure :
 - erreurs typographiques ;
 - erreurs cognitives ;
 - erreurs phonétiques.

Ces erreurs sont principalement liées à la saisie de données manuscrites, à des fautes de frappes, liées à des touches voisines sur un clavier ou encore la transcription téléphonique de données.

4.4.3. Techniques de comparaison

La comparaison de deux enregistrements, sans clé commune doit passer par deux étapes :

- une étape de comparaison de tous les champs communs entre les deux enregistrements ;
- une étape de compilation et d'analyse des résultats de comparaison pour la prise de décision sur le rapprochement.

La première étape concentre la majorité des travaux nécessaire au rapprochement, elle influencera largement les résultats issus de la deuxième étape.

Cette partie se focalisera plutôt sur les techniques permettant la comparaison champ à champ.

4.4.3.1. Méthodes empiriques

La façon la plus simple de mesurer la similarité entre deux champs est alors d'utiliser un opérateur « = » avec une réponse binaire si le champ est strictement égal ou non. Lorsque les champs à comparer sont de type valeur entière, ou une date, cette méthode peut avoir sa pertinence.

Néanmoins il est possible de définir d'autres méthodes plus robustes permettant la comparaison de deux valeurs entières s_1 et s_2 :

- en quantifiant la différence relative entre les deux valeurs:

$$\sigma(s_1, s_2) = 1 - \lambda \left(\frac{|s_1 - s_2|}{\max(s_1, s_2)} \right)$$

- si ces enregistrements sont bornés dans $[a, b]$ il est possible de préciser la distance relative à l'intervalle :

$$\sigma(s_1, s_2) \mid s_1, s_2 \in [a, b] = 1 - \lambda \left(\frac{|s_1 - s_2|}{b - a} \right)$$

Le paramètre λ (> 0) permet de pondérer la différence entre les deux valeurs suivant les champs à comparer, si la valeur est négative, on ramène la probabilité à 0.

Ces méthodes peuvent alors s'adapter à toute donnée contenant des valeurs numériques, comme les numéros de rue, codes postaux, âges ou dates.

Cependant, ces méthodes ne sont pas toujours adaptées en tant que telles à la plupart des champs. Afin de prendre en compte au mieux les erreurs humaines, les chiffres formant les nombres doivent être analysés un à un, et non de façon globale : les deux numéros de rue 99 et 100 sont bien plus éloignés dans la réalité que 99 et 69 par exemple.

Cette méthode manquant cruellement de souplesse, d'autres, plus étoffées sont apparues qui consistent à mesurer la similarité entre deux valeurs, le plus souvent estimé sous forme d'une probabilité ou d'un score.

4.4.3.2. Méthodes évoluées de comparaisons de chaînes de caractères

Damereau [182] puis Peterson [183] ont montré que plus de 80% des erreurs de transcription de données se limitaient à une de ces quatre catégories :

- omission d'une lettre ;
- substitution d'une lettre ;
- insertion d'une lettre ;
- échange de deux lettres adjacentes ;

Concernant les noms et prénoms de personnes, on peut se trouver face à deux autres catégories [178] :

- l'écriture des marques de ponctuation dans les noms de famille : apostrophes, espaces, tirets ou accents ;
- la confusion entre le nom de famille et le nom de jeune fille pour les femmes mariées.

Par la suite, Kukich [184] a aussi mis en évidence la probabilité supérieure que les lettres substituées ou insérées étaient des caractères voisins sur le clavier, une substitution [r/t] est plus probable que [r/p]. L'étude menée par Pollock [185] a montré que plus de 90% des mots mal orthographiés ne présentaient qu'une seule erreur. La situation de l'erreur dans le mot se présente le plus souvent dans sa troisième partie et uniquement dans de rares cas en première lettre.

De ces considérations sont nées des méthodes qui développent une certaine robustesse vis-à-vis de ces différents types d'erreurs. Ces algorithmes calculent alors une probabilité, estimant la *distance* qui sépare les deux chaînes comparées. Ensuite, le rapprochement probabiliste nécessite aussi la définition d'un seuil qui peut être adapté en fonction des circonstances. En dessous de ce seuil les deux valeurs seront considérées comme différentes et au dessus égales.

Plusieurs études : [186-190] se sont intéressées aux méthodes de comparaison de chaînes de caractères. On distingue ainsi deux grandes familles :

- les algorithmes de mesure de similarité appelés aussi « pattern matching » ;
- les algorithmes phonétiques, s'appuyant sur la prononciation des mots.

4.4.3.3. Principaux algorithmes de « pattern matching »

L'origine de ces algorithmes tient à la détection et corrections d'erreurs nécessaires à assurer la qualité d'une télécommunication. Par la suite, l'obligation décennale du recensement de la population américaine a vu sa complexité augmenter exponentiellement. Ainsi, la distance de Hamming [191], en 1950 influença Levensthein qui a conçu l'une des premières méthodes de mesure de similarité entre deux chaînes de caractères, depuis largement utilisée au sein des correcteurs orthographiques. Par la suite, Peterson, lors de son étude sur les erreurs de frappe dans les systèmes hospitaliers [183] a proposé l'algorithme LCS « Longest Common Substring ». Vient ensuite la distance de Jaro [192] puis l'amélioration par Winkler [186], pour constituer la distance de Jaro-Winkler dont la qualité lui a permis d'être repris dans l'implémentation de la méthode **strcmp()** en langage C.

Distance de Levensthein

La mesure de cette distance, dans l'algorithme original consiste simplement à comptabiliser, entre deux chaînes données en entrée, le nombre de permutations, substitutions, insertions et suppressions de lettres. Plus cette distance est élevée, plus les deux chaînes sont différentes. L'algorithme original n'effectuant qu'un compte du nombre de permutations, il est nécessaire de le normaliser pour obtenir une probabilité. Le nombre maximal de permutations possible étant égal à la taille maximale de la chaîne en entrée, la normalisation consiste juste à appliquer, partant de deux chaînes S_1 et S_2 :

$$p(S_1, S_2) = 1 - \left(\frac{\text{Levensthein}(S_1, S_2)}{\max(|S_1|, |S_2|)} \right)$$

```

Nom: Distance de Levensthein
Rôle: Mesure la distance entre deux chaînes de caractères
Entrée:  $S_1$  : Chaîne de caractères,  $S_2$  : Chaîne de caractères
Sortie: R : Entier
Déclaration: D : Entier[ $S_1$ .longueur(),  $S_2$ .longueur()]

début
  pour i ← 0 à  $S_1$ .longueur() faire
    D[i,0] ← i
  finpour
  pour j ← 0 à  $S_2$ .longueur() faire
    D[0,j] ← j
  finpour

  pour j ← 1 à  $S_2$ .longueur() faire
    pour i ← 1 à  $S_1$ .longueur() faire
      si  $S_1[i] = S_2[j]$  alors
        D[i,j] ← D[i-1,j-1]
      sinon
        D[i,j] ← min(D[i-1,j], D[i,j-1], D[i-1,j-1])
      finsi
    finpour
  finpour
  R ← D[ $S_1$ .longueur(),  $S_2$ .longueur()]
fin
  
```

Figure 57 - Algorithme de calcul de la distance de Levensthein

Algorithme LCS « Longest Common Substring »

Le problème de trouver la plus grande sous-chaîne commune à deux chaînes est un grand classique de l'algorithmique. Peterson [183], a eu l'idée, afin de comparer deux chaînes de caractères, de mesurer non pas la similarité mais la divergence. La méthode est simple, il suffit pour cela de supprimer de ces deux chaînes les sous-chaînes les plus grandes, jusqu'à ce que la plus grande chaîne commune soit un caractère.

Par exemple, comparons *maréchal* et *marchandage* :

$$\text{LCS}(\text{maréchal}, \text{marchandage}) = \text{mar}$$

Donc on retire aux deux chaînes la sous-chaîne *mar* :

$$\text{LCS}(\text{échal}, \text{chandage}) = \text{cha}$$

On obtient donc *él* et *ndage*. Le calcul de distance peut alors être calculé en divisant la taille totale des deux chaînes résultat et en les divisant par le min, max ou taille moyenne des deux chaînes

de départ. Ici, on obtient $\frac{7}{9}$, $\frac{7}{11}$ ou $\frac{7}{10}$ respectivement avec le min, max et moyenne. Pour une mesure de probabilité entre deux chaînes S_1 et S_2 : on peut la calculer à l'aide du maximum de longueur des deux chaînes.

$$1 - \left(\frac{|LCS(S_1, S_2)|}{\max(|S_1|, |S_2|)} \right)$$

Cet algorithme est très performant pour le cas où les champs à comparer contiennent nom et prénom qui peuvent être inversés : les deux chaînes « *Paul De Vlieger* » et « *De Vlieger Paul* » ont une distance égale à 0.

Distance de Jaro et de Jaro-Winkler

Jaro [192], se charge de compter, entre deux chaînes S_1 et S_2 le nombre C de caractères communs et le nombre de permutations P acceptables dans la moitié de la longueur minimum des deux chaînes. La mesure de similarité se calcule alors ainsi :

$$Jaro(S_1, S_2) = \frac{1}{3} \left(\frac{C}{|S_1|} + \frac{C}{|S_2|} + \frac{C - P}{C} \right)$$

L'amélioration par Winkler [186], a simplement pris en compte les conclusions de Pollock [185] sur le fait que les erreurs n'arrivaient que très rarement en première partie du mot.

Nom: Distance de Jaro
Rôle: Mesure la distance entre deux chaînes de caractères
Entrée: S_1 : Chaîne de caractères, S_2 : Chaîne de caractères
Sortie: R : Réel
Déclaration: SC_1 : Chaîne de caractères
 SC_2 : Chaîne de caractères
 t : Entier

début
 $SC_1 \leftarrow \text{CaracteresCommuns}(S_1, S_2)$
 $SC_2 \leftarrow \text{CaracteresCommuns}(S_2, S_1)$
 $t \leftarrow 0$
pour $i \leftarrow 0$ à $SC_1.\text{longueur}()$ **faire**
 si $SC_1[i] \neq SC_2[i]$ **alors**
 $t \leftarrow t + 0.5$
 fin si
fin pour
 $R \leftarrow SC_1.\text{longueur}() / S_1.\text{longueur}() +$
 $SC_2.\text{longueur}() / S_2.\text{longueur}() +$
 $(SC_1.\text{longueur}() - t / SC_1.\text{longueur}()) / 3$
fin

Figure 58 - Algorithme de calcul de la distance de Jaro

Ainsi la distance de Jaro-Winkler se calcule en prenant en compte le nombre N de caractères communs en début des deux chaînes :

$$JaroWinkler(S_1, S_2) = Jaro(S_1, S_2) + \frac{N}{10} (1 - Jaro(S_1, S_2))$$

4.4.3.4. Principaux algorithmes phonétiques

Soundex

Malgré leur apparente complexité, les méthodes phonétiques sont les premières à être apparues pour comparer des chaînes de caractères. Ainsi, Soundex, breveté en 1918 par Russell [193] est la première méthode de comparaison phonétique de mots. Elle s'appuie sur l'origine physique de la formation des sons par l'être humain. Ce procédé, présenté d'un point de vue algorithmique en 1968 par D.Knuth [194] et réellement exploité par Hermansen en 1985 [195] consiste à transformer tout mot en un code sonore suivant la table de correspondance montrée en [Figure 59]. Le principe est simple, pour le mot « Catherine » on garde la première lettre « C » à laquelle on ajoute trois digits suivant la table de correspondance, les lettres absentes de la table sont ignorées et si la chaîne est trop courte on complète par des 0. On obtient alors « C365 ».

Class	1	2	3	4	5	6
Letters	B F P V	C G J K Q S X Z	D T	L	M N	R

Figure 59 - Table de correspondance lettres-code soundex en anglais

Ainsi, les déclinaisons « Cathrina », « Catarina » ou « Catarinella » auront le même code et seront considérées comme *phonétiquement* équivalentes.

Les limites de l'algorithme interviennent pourtant assez rapidement car pour des mots assez éloignés phonétiquement, on peut obtenir le même code. Ainsi « Citron » ou « Cadran » ont aussi pour code C365. De plus, selon Patman [196], de nombreuses réserves sont émises sur la qualité de Soundex. Il rapporte que :

- Soundex est trop dépendant de la première lettre ;
- la taille des chaînes à comparer doit être raisonnable si on se limite à un code à 4 digits ;
- certains « sons » composés de plusieurs consonnes ne sont pas pris en compte ;
- les champs composés de plusieurs mots ne prennent pas en compte les permutations.

La table de correspondance est aussi à adapter en fonction de la langue d'entrée utilisée. Ainsi, la [Figure 60] représente l'adaptation française de l'algorithme.

Classe	1	2	3	4	5	6	7	8	9
Lettres	B P	C K Q	D T	L	M N	R	G J	X Z S	F V

Figure 60 - Soundex en français

Phonex

Devant les lacunes de Soundex, des versions améliorées sont apparues, notamment Phonex [197], proposé par Lait en 1996. Le principe de Phonex est de transformer au préalable d'un codage Soundex les chaînes de caractères à traiter pour prendre en compte un maximum de subtilités du langage.

Il prend notamment en compte une transcription des sons qui s'expriment avec plusieurs consonnes, une des faiblesses de Soundex. Le contexte de chaque mot a aussi son importance, afin de coller au plus proche de la réelle prononciation. La lettre h par exemple, oubliée de Soundex car jugée non pertinente, a maintenant son importance lorsque « p » la précède par exemple.

Une version de Phonex a été adaptée à la langue française par Brouard [198], en prenant un maximum de subtilités de la langue. L'algorithme, plutôt que de calculer un code court transforme toute la chaîne en un nombre en base 22 exprimé en virgule flottante. Une explication complète du fonctionnement de Phonex est fournie en [Annexe 2].

4.4.3.5. Implémentation et test des algorithmes

De nombreuses autres méthodes de comparaisons de chaînes de caractères existent mais seules les plus efficaces selon Christen [189] ont été retenues ici, voir [Annexe 5].

Il apparaît que, pour les algorithmes de mesure de similarité ou de « pattern matching », LCS, Jaro et Winkler obtiennent les meilleurs résultats suivant si l'on compare les noms ou prénoms. Concernant les algorithmes phonétiques, seul Phonex tire vraiment son épingle du jeu, peu importe les données comparées.

Le mode opératoire proposé par Christen est, à partir d'un jeu de données personnelles contenant noms et prénoms, d'introduire de manière stochastique des biais parmi les champs. Ces biais pouvaient être des ajouts/suppressions de N caractères ou inversion de caractères. Parmi les tests effectués, on constate que la qualité des algorithmes de mesure de similarité est bien supérieure à celle des algorithmes phonétiques.

Cependant, le manque d'application sur des cas réels, comme deux bases de données comprenant des personnes communes ne permet pas de vérifier les résultats.

Simmetrics

Simmetrics [199] est une bibliothèque de fonctions de comparaison de chaînes de caractères. Elle est écrite en Java et diffusée en open source. Elle implémente notamment Jaro, Winkler, Soundex, Levensthein et en propose bien d'autres. L'avantage de cette bibliothèque est aussi dans sa facilité à implémenter d'autres méthodes, qui peuvent alors réutiliser tout le cadre pratique proposé par Simmetrics. Elle sera réutilisée pour tester ces différentes méthodes sur les données issues du projet RSCA. Elle permettra aussi d'implémenter d'autres méthodes en utilisant son environnement, comme Phonex-fr, qui n'existe pas dans cette bibliothèque.

4.4.4. Medical Data Linkage

Les premières mentions de « medical record linkage » dataient déjà de 1969 par Acheson [181]. Depuis, les principaux auteurs de méthodes de comparaison de chaînes de caractères Levensthein voir [4.4.3.3], Jaro [200] puis Winkler [186] se sont naturellement intéressés au contexte médical et proposé des solutions.

Depuis, des systèmes complets se sont spécialisés dans le « medical record linkage », [201-203], ou plus récemment en open-source [204] avec interface graphique.

Des études similaires ont aussi été menées en France dans le cadre du centre hospitalier de Mulhouse [205, 206], proposant ainsi une méthode rudimentaire mais néanmoins efficace de rapprochement d'identités.

4.4.4.1. Problème théorique de rapprochement d'identités

Le rapprochement d'une identité peut rencontrer divers cas de figure, suivant si le rapprochement est correct ou non. On peut distinguer 4 cas différents :

- les vrais positifs : lorsque le rapprochement s'est effectué correctement (même patient) ;
- les vrais négatifs : lorsque le rapprochement ne s'est pas effectué (patients différents) ;
- les faux négatifs : lorsque le rapprochement ne s'est pas effectué alors que les patients comparés étaient les mêmes. Dans ce cas un patient aura deux identifiants ;
- les faux positifs : lorsque le rapprochement s'est effectué alors que les patients sont différents.

Bien entendu ce dernier cas est à éviter à tout prix car il signifierait que deux patients auraient le même identifiant, ce qui peut amener à des erreurs graves.

Du point de vue théorique, la distribution de la comparaison des patients ressemblerait à la [Figure 61]. La distribution des vrais négatifs (non rapprochement) et vrais positifs (rapprochement)

suivent deux gaussiennes qui se chevauchent. Cette partie en recouvrement fait apparaître les deux autres cas de figure : les faux positifs et faux négatifs.

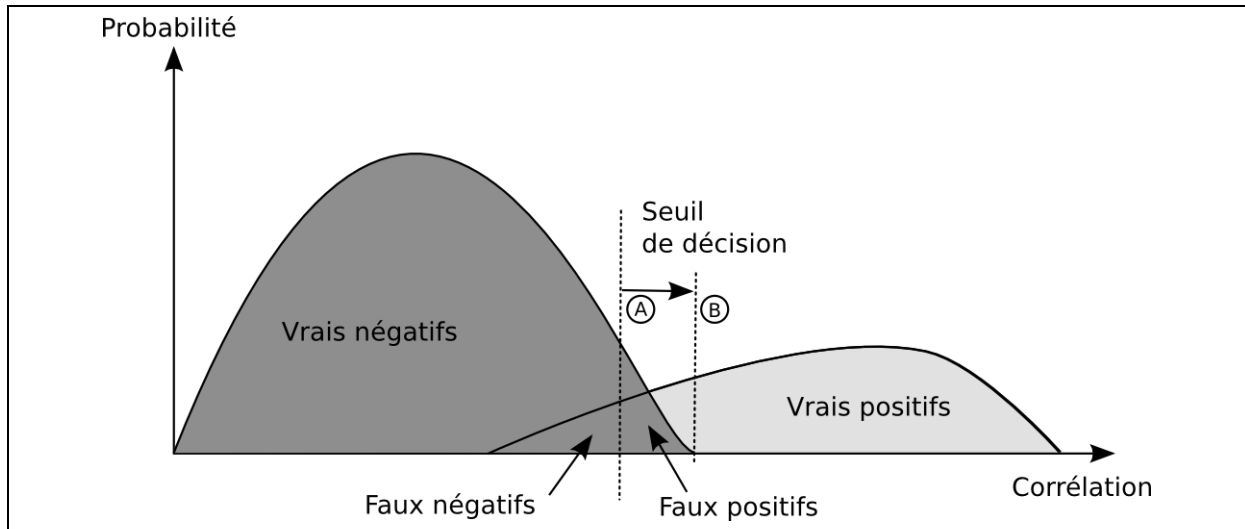


Figure 61 - Problème théorique de rapprochement d'identités

Dans ce cas de figure la décision de rapprochement des identités est assez simple, vu qu'il est indispensable de minimiser les faux positifs, le seuil de décision est à fixer en ②. Dans ce cas la proportion de faux négatifs ne serait pas minimale mais l'objectif est de minimiser les faux positifs.

4.4.4.2. Problème pratique de rapprochement d'identités

En pratique, lorsque les biais entre deux patients est important, la zone de recouvrement des faux positifs et faux négatifs est plus large, l'application d'un seul seuil ② ne permettrait pas d'avoir un taux de rapprochement suffisamment élevé pour un système performant de rapprochement d'identités. Comme montré en [Figure 62], la création de deux seuils ① et ② permettrait de définir, entre deux zones de décision automatique, une zone intermédiaire ou l'intervention d'une personne dédiée permettrait d'effectuer le choix du rapprochement, ou non des deux patients.

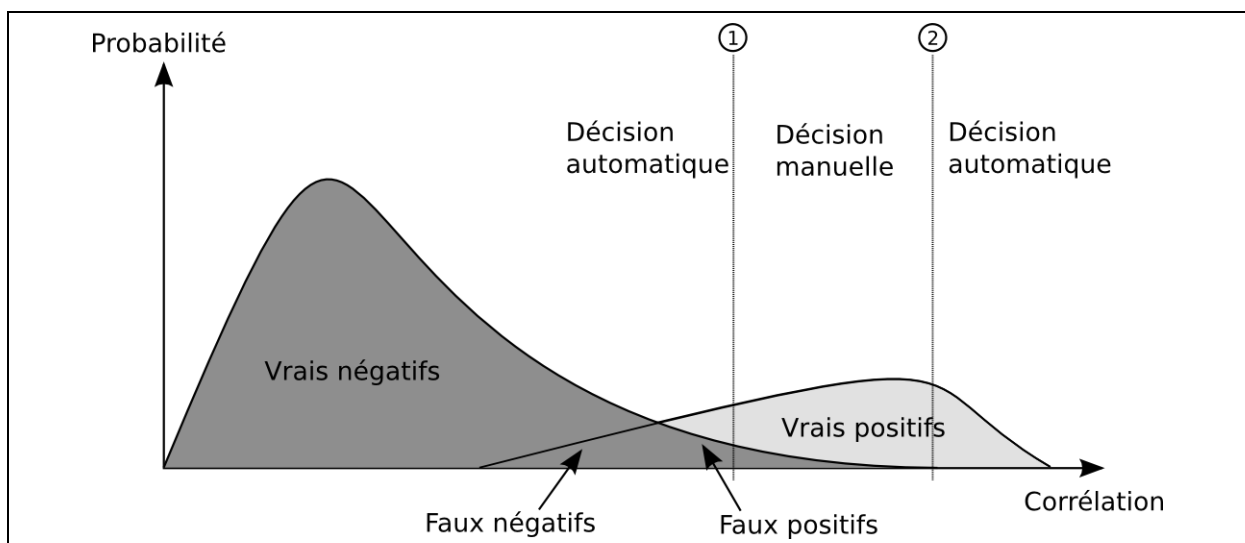


Figure 62 - Problème pratique de rapprochement d'identités

4.4.4.3. Spécificité de RSCA (grille de données patient distribuées)

Outre le « medical record linkage », le réseau RSCA pose un problème plus conséquent de la distribution des données. En effet, là où les différents systèmes ne se chargent pas de comparer des sources de données locales, il est nécessaire ici de proposer un modèle respectant les contraintes de la répartition de l'information.

Un réseau de grille, comme tout réseau, ne peut pas garantir une fiabilité et une qualité de service nécessaire à la comparaison de sources de données en temps réel. Pour cela, les différentes requêtes de comparaisons de patient devront se faire majoritairement de façon asynchrone. Cette méthode, bien que présentant l'avantage d'être robuste vis-à-vis de la qualité lien réseau a l'inconvénient de prendre un temps supérieur et surtout indéterminé lors de l'exécution.

Processus de comparaison spécifique à RSCA

Le processus d'insertion d'une source de donnée dans RSCA qui a été adopté, représenté en [Figure 63], consiste en plusieurs étapes. D'abord, à partir d'une source de donnée A, une étape de standardisation puis de fusion et d'indexation pour permettre la comparaison champ à champ avec les autres bases de données est effectuée. Cette phase de comparaison est la plus importante et va aboutir à trois possibilités de décision :

- si le score de rapprochement est supérieur au seuil ② alors il y a correspondance, le patient correspond déjà à un patient existant ;
- si le score de rapprochement est inférieur au seuil ① alors il y a non correspondance, un nouvel identifiant patient est alors créé ;
- si le score se situe entre les seuils ① et ② alors une intervention manuelle est nécessaire pour effectuer le bon choix, un choix automatique présenterait trop de risque de faux positifs.

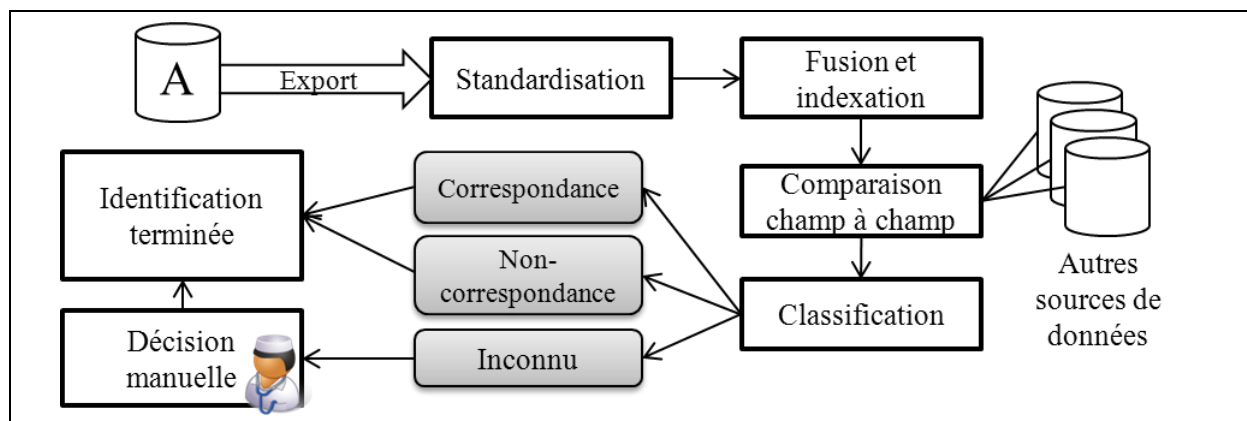


Figure 63 - Processus d'identification

4.4.4.4. Compérateur « champ à champ »

La pierre angulaire du projet étant dans cette problématique, ses performances influenceront beaucoup le fonctionnement du réseau. Vu qu'il n'existe pas de standard de performance dans ce domaine, encore moins en langue française, il est nécessaire de proposer une méthode la plus flexible et évolutive possible afin de tester plusieurs méthodes de rapprochement.

Choix de l'architecture de comparaison de champs

Afin de garantir cette flexibilité maximale dans les méthodes de comparaison de chaînes, l'architecture, montrée en [Figure 64] adopte le modèle de conception « Strategy » [207], complètement adapté à la situation. Compte tenu de deux sources de données, une classe principale « Comparator » associée à chaque champ à comparer depuis ces deux sources un ensemble de règles régies par une implémentation d'une stratégie particulière.

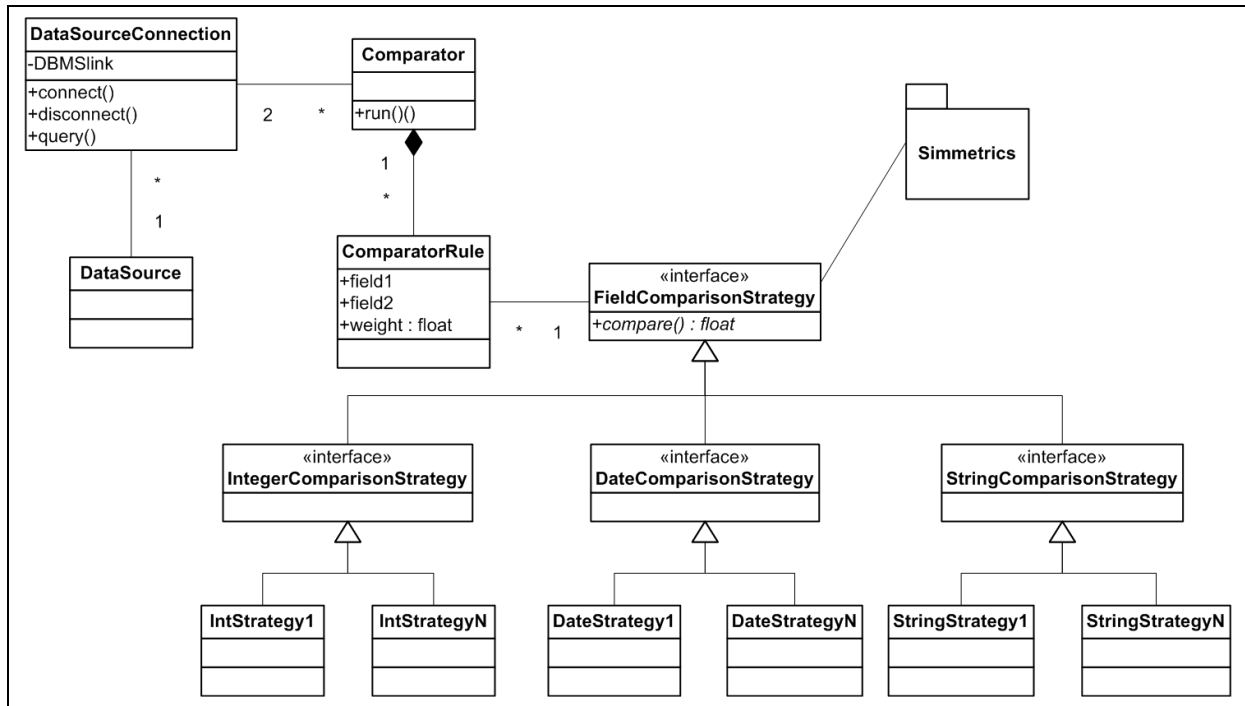


Figure 64 - Diagramme de classes - Comparateur de champs

Ces stratégies sont séparées en trois classes, suivant le type de données à comparer « Entier », « Date » ou « Chaîne de caractères ». L'implémentation réelle de ces stratégies peut alors s'appuyer sur la bibliothèque externe Simmetrics pour réaliser la fonction de comparaison.

Afin d'augmenter le niveau de flexibilité, une stratégie de comparaison peut aussi s'appuyer sur une autre. Par exemple, la stratégie de comparaison d'une date peut s'appuyer sur trois comparaisons successives des composants jour, mois, année qui composent la date issue de la stratégie « Entier ».

4.4.4.5. Généralisation au « Record Linkage »

La partie précédente sur la comparaison champ à champ ne suffit pas à créer un véritable système de « Record Linkage ». La généralisation doit permettre de décider, en comparant tous les champs équivalents entre deux sources de données si oui ou non deux enregistrements sont similaires.

Classes de décision

Lors de la comparaison de deux champs, on peut définir plusieurs critères sur la décision. En effet, la décision à prendre n'est pas la même si le nom de famille est différent ou si l'adresse est différente. Pour cela quatre paramètres sont définis :

- poids (correspondance) : définit le poids que représente la correspondance de deux champs ;
- poids (non-correspondance) : définit le poids que représente la non-correspondance de deux champs ;
- blocage: attribut qui définit qu'en cas de non-correspondance le processus de comparaison est interrompu.

L'ensemble de ces éléments permet de calculer le score global de rapprochement entre les deux enregistrements.

4.5. PRESENTATION DES RESULTATS

Cette partie présente les résultats préliminaires d'identification et de rapprochement d'identités patient. Le but étant de vérifier les forces et faiblesses de ces différentes méthodes.

4.5.1. *Expérimentation sur données simulées*

Protocole expérimental

Afin de mener à bien les tests sur le rapprochement d'identité, plusieurs prérequis sont nécessaires.

Jeu de données utilisé

Devant la haute sensibilité des données utilisées dans RSCA, il n'est pas possible, pour des raisons de respect de la vie privée de les utiliser pour le test des méthodes de « data linkage ».

Une des sources de données « réelle » et exploitable trouvée en libre accès est la base « Social Security Death Index », qui est une base de données des personnes décédées aux USA [208].

Le protocole de comparaison consiste alors à générer une base de données suffisamment grande, puis d'en faire une version dupliquée avec introduction de biais stochastiques dans les noms, adresses etc. Les champs disponibles dans la base de données SSDI sont :

- Social security number (clé) ;
- Nom, prénom ;
- Date de naissance ;
- Adresse (code + ville).

Introduction des biais

Deux jeux de données ont tout d'abord été créés, en contenant chacun 10.000 enregistrements. Une proportion de 12% d'enregistrements est commune aux deux bases de données.

Parmi ces enregistrements, différents types de biais ont été insérés :

- suppression / ajout de caractère ;
- inversion de caractères ;
- substitution de caractères.

La proportion de biais introduits dans les champs est de 1 pour 1, c'est-à-dire pour N enregistrements, N biais sont introduits de façon aléatoire. On obtiendra une répartition suivant une loi de poisson de paramètre $\lambda = 1$ qui suit au plus près les observations de Damerau [182].

Les positions des caractères sur lesquels les biais vont être introduits seront quant à elles choisies de façon totalement aléatoire et indépendante.

Configuration des paramètres

Les paramètres utilisés pour la comparaison champ à champ sont montrés en [Figure 65].

	Nom	Prénom	Date de naissance	Adresse
Type	String	String	Date	String
Précision	•••	•••	•••	•
Élément bloquant	X		X	
Poids (similaire)	•••	•••	•••	•
Poids (différent)	•••	•	••	•

Figure 65 - Paramètres utilisés pour la comparaison

Le paramètre « Élément bloquant » permet, en cas de non-correspondance, de refuser catégoriquement le rapprochement des identités. Cette mesure permet, en cas d'homonymes, de ne pas rapprocher les identités si la date de naissance est radicalement différente.

Paramètres des algorithmes de comparaison

Dans cette étude, les algorithmes choisis pour mesurer leur performance sont Jaro-Winkler et Soundex en version anglophone. Les seuils de rapprochement ont été fixés à 0.9. Cette procédure permet de vérifier la qualité des différentes méthodes de comparaison de chaînes.

Résultats bruts

Cette partie présente les résultats de façon brute, en indiquant, pour chaque méthode et pour chaque champ, le taux de vrais positifs (VP), faux négatifs (FN) et faux positifs (FP), représentés en [Figure 66]. La précision se calcule de la manière suivante : $Précision = \frac{VP}{VP+FN+FP}$

Champ - Méthode	VP	FN	FP	Résultat	Précision
Nom – Jaro-Winkler	11.53	1.21	0.06	96.08	90.08
Nom – Soundex-US	9.33	1.14	0.11	77.75	88.19
Prénom – Jaro-Winkler	13.11	2.21	0.09	109.25	85.07
Prénom – Soundex-US	10.37	1.93	0.13	86.42	83.43
Adresse – Jaro-Winkler	9.82	1.72	0.11	81.83	84.29
Adresse – Soundex-US	7.41	1.72	0.19	61.75	79.51

Figure 66 - Résultats bruts de data linkage

Ces résultats confirment les performances des différents algorithmes, avec un net avantage pour Jaro-Winkler. De réelles difficultés sont rencontrées sur les adresses, avec un taux de rapprochement bien en deçà des noms ou prénoms. L'adresse contenant une succession de lettres et de chiffres, il n'est pas évident de mesurer la distance entre deux adresses, même très légèrement différentes.

Résultats complets de rapprochement d'identités

Toujours sur le même jeu de données, mais en prenant en compte cette fois-ci la configuration proposée en [Figure 65], le rapprochement d'identité est alors plus précis, comme montré en [Figure 67].

Nom + Prénom + Adresse	VP	FN	FP	Résultat	Précision
Jaro-Winkler	11.63	0.24	0.01	96.91	97.8
Soundex-US	9.93	1.08	0.04	82.75	92.08

Figure 67 - Résultats combinés de data linkage

Compte tenu des biais introduits dans les deux jeux de données, les résultats sont plutôt bons dans l'ensemble, avec un taux de résultats de 96.91% pour Jaro-Winkler. Le principe de bloquer le rapprochement d'identité si le Nom présente un résultat inférieur à un seuil permet d'améliorer significativement le taux de faux positifs mais au détriment des vrais positifs.

Cependant, aucune des deux méthodes ne permet d'éliminer complètement les faux positifs.

Améliorations possibles

Dans l'idée d'améliorer encore le rapprochement d'identités, dans le cas où les données sont réelles, c'est-à-dire avec un taux d'erreurs bien inférieur à une distribution poissonnienne, les taux nominaux de rapprochement seront normalement plus élevés.

Une autre amélioration possible est de détecter plus précisément des seuils de rapprochement à adopter pour maximiser la précision. C'est l'objet de l'étude suivante.

4.5.2. Expérimentation sur données réelles

Un jeu de données semblables nous a été fourni dans le cadre du projet RSCA contenant un échantillon de 70000 enregistrements conformes à la population locale au projet, noté A.

Protocole expérimental

Un protocole similaire à l'expérimentation sur données simulées a été utilisé : une copie A' de la base de données a été créée contenant un ensemble de biais suivant une distribution poissonnienne. Le jeu de données ainsi obtenu sera assez fortement modifié, certaines modifications pouvant être très préjudiciables aux algorithmes de comparaison comme la modification de la première lettre pour les algorithmes phonétiques.

La comparaison utilise successivement l'algorithme de Jaro-Winkler et Phonex, cette fois-ci en version francisée compte tenu du jeu de donnée à disposition voir [Annexe 3].

Dans un premier temps, seuls les deux champs nom et prénom ont été utilisés dans l'expérimentation. Pour chaque enregistrement de la base A', celui-ci est comparé à l'ensemble de la base A. Seul le meilleur enregistrement est gardé comme correspondant.

Objectif

Cette expérimentation aura pour objectif de déterminer les valeurs des seuils à adopter pour maximiser les performances du rapprochement d'identités et visualiser le comportement du rapprochement dans cet environnement.

Résultats

La comparaison effectuée, on retrouve, comme montré en [Figure 68] le rapport entre proportion de vrais positifs, faux négatifs et faux positifs en fonction du seuil de décision issu des paramètres de l'algorithme de rapprochement d'identités. L'indice F, proposé par Christen [189], qui correspond à :

$$F = 2 \left(\frac{P \cdot R}{P + R} \right) \text{ avec } P = \frac{VP}{VP + FP} \text{ et } R = \frac{VP}{VP + FN}$$

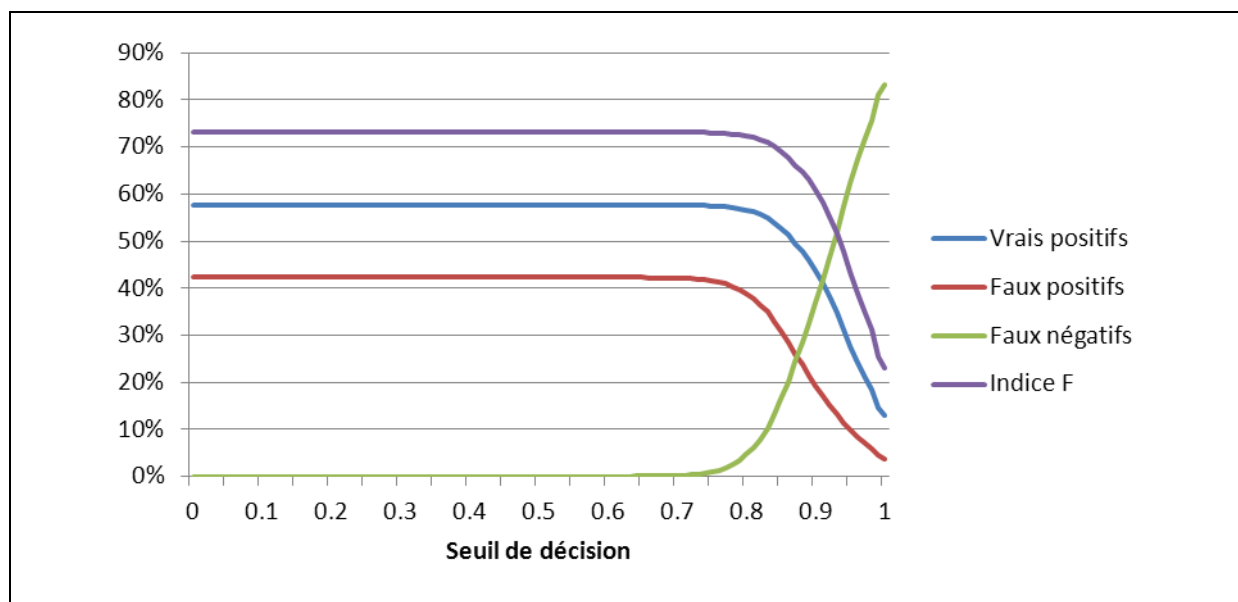


Figure 68 - Résultat de comparaison en fonction du seuil

Analyse

Ces résultats ne sont clairement pas déterminants quant aux performances de la méthode. Après étude approfondie des résultats, on constate que le jeu de données biaisé l'est souvent trop fortement par rapport à la robustesse de JaroWinkler et surtout de Phonex-fr. Ensuite, les seuls noms et prénoms ne suffisent souvent pas à prendre une décision, les confusions sur les homonymes sont très nombreuses.

La seule conclusion tangible se trouve au niveau du seuil de décision qui se situe entre 0.8 pour une limite basse et 0.95 pour la limite haute.

4.5.3. Expérimentation étendue sur données réelles

Compte tenu de l'analyse précédente, une nouvelle expérimentation a été menée, cette fois-ci en prenant en compte la date de naissance et l'adresse complète de la personne. Le jeu de données simulé à partir des bases des associations de dépistage (70000 individus) reprend les mêmes caractéristiques et les paramètres de comparaison sont identiques à l'expérimentation précédente.

Résultats

La [Figure 69] présente les résultats de cette expérimentation. On constate clairement cette fois-ci que les performances sont nettement meilleures, avec un taux nominal de moins de 6% de faux positifs.

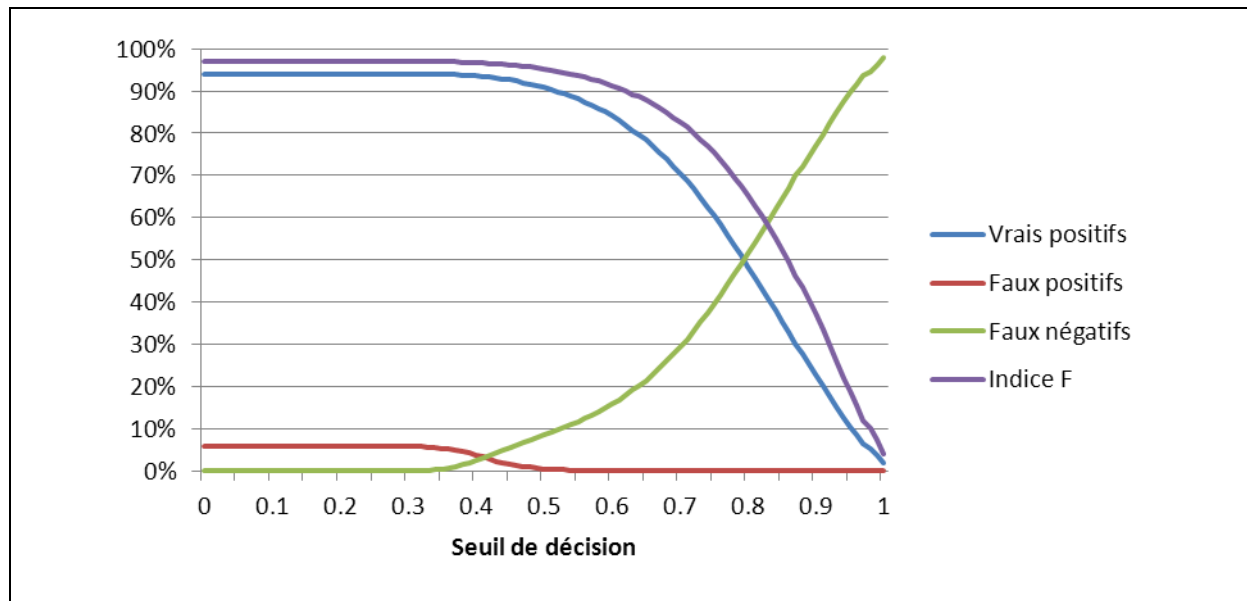


Figure 69 - Résultats de comparaison - Date de naissance et adresse incluse

Analyse

Dans ce contexte, le seuil de décision peut être fixé en limite basse à 0.4 et en limite haute à 0.7 car en deçà, plus aucun faux positif n'apparaît. Cette observation est lourde d'enseignements : le nombre de champs disponibles à la comparaison est déterminant sur les performances de rapprochement d'identités. Même peu significatifs, les champs comme l'adresse ou le nom de médecin prescripteur peuvent influencer les résultats.

Limitations

L'algorithme de comparaison a pourtant un inconvénient majeur, sa vitesse d'exécution. Le fonctionnement de celui-ci consiste à comparer champ à champ tous les enregistrements disponibles. Sur la base de test qui nous a été fournie, de 70000 individus, il faut plus de 10 heures pour calculer l'ensemble des rapprochements à une moyenne de 2 par seconde.

Bien que l'algorithme ne soit pas optimisé, la complexité de celui-ci en $O(n^2)$ explique ses performances : comme montré en [Figure 70], la comparaison consiste en quatre boucles imbriquées :

- la première parcourt tous les enregistrements du premier jeu de données ;
- la deuxième parcourt tous les enregistrements du deuxième jeu de données ;
- la troisième liste tous les champs des enregistrements ;
- la dernière liste les algorithmes de comparaison : JaroWinkler et Phonex-fr.

Ainsi, pour un jeu de données de 70.000 enregistrements qui comportent quatre champs, chaque instance de l'algorithme de comparaison est exécuté $70.000 \times 70.000 \times 4$, soit 19,6 milliards de fois.

```

Nom: Algorithme de matching
Rôle: Permet de rapprocher les identités des deux jeux de données en entrée
Entrée:  $S_1$  : Jeu de données,  $S_2$  : Jeu de données
Champs : Tableau[Champ]
Algos : Tableau[Algorithmes de comparaison]
Sortie: Matching : Tableau[Enregistrement]
Déclaration: resultat : Float, max : Float, maxE : Enregistrement

```

```

début
  pour i ← 0 à D1.longueur() faire
    max ← 0
    pour j ← 0 à D2.longueur() faire
      resultat ← 0
      pour k ← 0 à Champs.longueur() faire
        pour l ← 0 à Algos.longueur() faire
          resultat ← resultat + Algos[l].calc(D1[i].champ(k), D2[j].champ(k))
        finpour
      finpour
      resultat ← resultat / (Algos.longueur() * Champs.longueur())
      si resultat > max alors
        max ← resultat
        maxE ← D2[j]
      finsi
    finpour
    Matching[i] ← maxE
  finpour
fin

```

Figure 70 - Algorithme simplifié de matching

Perspectives et améliorations

Un certain nombre d'améliorations sont possibles pour accélérer l'exécution de l'algorithme. Car là où celui-ci s'exécute en une dizaine d'heures sur la base de test de 70000 enregistrements, la complexité quadratique ferait exploser le temps d'exécution si le nombre d'enregistrements s'agrandit trop.

Plusieurs pistes d'amélioration sont alors à envisager :

- arrêt du processus de comparaison si un champ clé est en dessous d'une certaine valeur : ex : si le nom est totalement différent, on passe alors à l'enregistrement suivant ;
- lancement de l'algorithme en multi processus : les comparaisons de la boucle principale sont indépendantes, elles peuvent donc être exécutées en parallèle ;
- amélioration des algorithmes de comparaison : JaroWinkler et Phonex-fr, une piste est fournie en [Annexe 6] avec une implémentation en utilisant les processeurs graphiques (GPGPU).

CONCLUSION ET PERSPECTIVES

L'ensemble des solutions proposées dans ce chapitre, liées à la sécurité des données médicales, à l'identification du patient et au rapprochement d'identités permet de proposer une méthode complète de gestion du patient dans un environnement distribué.

En premier lieu, le modèle de sécurité adopté respecte toutes les recommandations fixées par les lois françaises dans le domaine. En adoptant un chiffrement systématique de l'information et un double contrôle de l'authentification par CPS et de l'autorisation par VOMS, l'accès au réseau est strictement contrôlé et surtout traçable en cas d'utilisation frauduleuse.

Le modèle d'identification fourni permet de s'affranchir des obstacles que représente l'accès réparti à l'information médicale. Il propose en outre une flexibilité de gestion de l'identification qui permettra d'envisager une évolution à long terme, notamment lorsque l'INS sera déployé sur tout le territoire. Cette flexibilité permet aussi de suivre au plus près la vie des données médicales du patient, avec une évolution qui pourra le suivre tout au long de son parcours.

Le système de rapprochement d'identités, quant à lui permet, en garantissant la confidentialité et la sécurité du patient, de maintenir un haut niveau de cohérence tout au long du réseau. L'automatisation du processus permet de rapprocher une grande majorité de patients, avec un risque de fausse identification minime.

Cependant, certaines phases de l'algorithme de rapprochement doivent encore être validées, notamment au niveau des performances dans un environnement de production. L'état d'avancement du projet RSCA ne permet pas encore d'obtenir l'agrément pour utiliser les données en dehors du système de soin. Ainsi, un travail de validation et d'expérimentation sur des données réelles, sans partie simulée (au niveau des biais) est nécessaire. Cependant, compte tenu des performances obtenues lors de cette dernière expérimentation, les niveaux de rapprochement automatique de patient devraient être assez confortables pour n'avoir que très peu d'interventions humaines.

Dans un autre registre, le passage à l'échelle doit encore être vérifié, pour le système d'identification comme pour le rapprochement d'identités. A savoir quel sera leurs comportements lors d'utilisations intensives comme pour une enquête épidémiologique ou lors d'un ajout massif d'une grande quantité de données au réseau. Des mécanismes de gestion de la charge seront nécessaires pour assurer une qualité de service optimale.

Conclusion générale

Les travaux présentés dans cette thèse résument les trois années qui ont été nécessaires pour bâtir, sur les fondations offertes par les technologies des grilles informatiques le projet *Réseau Sentinelle Cancer Auvergne*. De nombreuses étapes ont été nécessaires pour la mise en œuvre technique de l'application. Le fruit des nombreuses discussions avec les personnels de santé, à l'origine du projet, a permis tout d'abord de mettre en place un cahier des charges fonctionnel.

Une longue étape d'analyse technique, technologique et légale s'en est suivie, pour étudier la faisabilité du projet et la compatibilité des outils issus des grilles informatiques maîtrisés par l'équipe PCSV du Laboratoire de Physique Corpusculaire de Clermont-Ferrand, co-hébergeant cette thèse avec l'équipe ERIM, devenue en 2011, ISIT¹.

Une étude approfondie a ensuite été menée afin de garantir la sécurité des données, l'authentification forte des utilisateurs et la confidentialité des patients. Celle-ci a amené le développement d'un modèle d'authentification utilisant les cartes de professionnel de santé (CPS) afin de le rendre compatible avec les couches de sécurité des grilles informatiques.

Conjointement, une demande d'autorisation à la CNIL a été déposée, ce qui a d'ailleurs marqué le lancement officiel du projet.

Par la suite, le rapprochement avec la société maat-G, impliquée dans de nombreux projets similaires de recherche a alors permis une description technique du projet et une mise en place du modèle de l'architecture du réseau en fournissant la pièce maîtresse qui équipe les serveurs dans chaque site, à savoir la Pandora Gateway.

L'infrastructure du projet s'est ainsi déployée, en équipant peu à peu les structures de dépistage des cancers en Auvergne et les laboratoires et services de pathologie partenaires du projet. Un premier prototype a alors été déployé, sur des données factices, afin de prouver que le concept de grille dédiée à l'échange de données médicales était viable.

Cependant, un ensemble de développements a été nécessaire afin de rendre parfaitement opérationnel le réseau. Il fallait, afin de doter l'infrastructure d'une bonne cohérence de l'information, un modèle d'identification qui s'affranchit des contraintes légales sur l'identité des patients en France et garantissant une confidentialité sans faille des données médicales. En l'absence d'une solution utilisable en France, il a fallu bâtir un nouveau modèle adapté au caractère distribué du réseau qui respecte l'ensemble des contraintes légales régissant le transport des données médicales. L'utilisation d'une fédération d'identifiants anonymes représentant chacun une petite partie des données médicales apporte une certaine souplesse d'utilisation. Le couplage de ce modèle

¹ Image Science for Interventionnal Techniques, www.u-clermont1.fr/isit

à un ensemble de méthodes de chiffrement asymétrique de l'information, proposée par l'infrastructure à clé publique (PKI) garantit la confidentialité des données du patient.

Ce modèle produit, une autre étape de rapprochement des patients a été nécessaire pour pouvoir mettre en relation les bases de données distribuées. Un ensemble de solutions ont été testées, issues des méthodes les plus évoluées de comparaison de chaînes de caractères. Finalement un algorithme de rapprochement s'appuyant sur l'utilisation conjointe d'une méthode phonétique et algébrique de comparaison est proposé. Une mesure de performances et de rapidité de l'algorithme de comparaison a aussi été fournie, avec quelques pistes pour en améliorer les performances.




Le domaine du « medical record linkage » présente de nombreuses perspectives de recherche car de nombreux travaux, surtout sur la prononciation française de mots restent encore à faire. L'environnement de recherche idéal que propose ce projet dans ce domaine, avec de véritables données d'une population ciblée sur un territoire délimité est une opportunité trop rare pour ne pas être exploitée...

Les difficultés rencontrées lors de la mise en œuvre du projet ont surtout été d'ordre légal, ce qui explique que ce point occupe une place non-négligeable dans ce mémoire. L'accord de la CNIL pour ce projet n'a pas été obtenu à ce jour et cela pour des raisons indéterminées. Le silence de la CNIL à ce sujet nous a obligé à étudier d'autres solutions pour permettre la mise en œuvre du réseau.

Le projet ANR GINSENG (Global Initiative for Sentinel E-health Network on Grid), démarré début 2011, a pour objectif de bâtir, sur la base de RSCA, une infrastructure capable de fédérer un ensemble de sources de données hétérogènes à des fins épidémiologiques. Ce projet a des ambitions très prometteuses sur le plan technique mais surtout sur le plan légal, en demandant une autorisation CNIL par le biais du CCITRS [9], présenté en partie [1.1.4].

L'avènement du *Réseau Sentinelle Cancer Auvergne* dépendra surtout du succès et de la capacité à convaincre du projet GINSENG.

Annexes

SIPATH - ANATOMIE ET CYTOLOGIE PATHOLOGIQUES <small>www.sipath.fr</small>			
 PARDIEU 18, av. Léonard de Vinci 63063 Clermont-Ferrand Cedex 1 Tél.: 04 73 28 51 70	RÉPUBLIQUE 105, av. de la République 63023 Clermont-Ferrand Cedex 2 Tél.: 04 73 99 46 00 Fax: 04 73 99 46 01	ROANNE 75, rue Général Giraud 42300 Roanne Tél.: 04 77 44 41 84 Fax: 04 77 72 33 51	VICTORIA 2, av. Victoria 03206 Vichy Cedex Tél.: 04 70 30 96 10 Fax: 04 70 98 27 42
<div style="display: flex; justify-content: space-between;"> <div>  </div> <div style="text-align: center;"> Suite de l'examen Concernant </div> </div>			
<p>CONCLUSION</p> <p>Tumorectomie QIE du sein droit : adénocarcinome infiltrant (canaire, à confirmer par immuno-histo-chimie) de grade 2 selon Scarff, Bloom et Richardson.</p> <p>La tumeur mesure 10 mm et son exérèse est totale avec des marges de 4 mm.</p> <p>Individualisation sur les coupes sériées du ganglion sentinelle d'un foyer de micro métastase de 1 mm sans rupture capsulaire, absente de la coupe vue en examen extemporané.</p> <p>PS : Rappel RO 80% 2+ RP 70% 2+ HER2 négatif</p> <div style="border: 1px solid black; padding: 5px; margin-top: 10px;"> <p>COMPTE RENDU COMPLEMENTAIRE</p> <p>La tumeur présente un marquage fort cytoplasmique par E Cadherine. Elle est donc bien d'origine canalaire.</p> </div> <p style="font-size: small; margin-top: 20px;"> <i>Etude immuno-histochimique : procédure XT.VIEW DAB V.3 Module Benchmark XT IHC Ventana.</i> <i>Références des anticorps utilisés en immuno-histo-chimie mammaire (Automate Benchmark VENTANA) : P.D. : ref 3591015, Microm, Clone SP1, RP : ref 799-2223, Ventana, Clone IE2, Her2 : ref 100-2996, Ventana, Clone 4B5, E cadherine : ref 180223, Zymed, Clone 4A2C7.K167(MIB1) : ref M7240, Microm, Clone MIB-1, Cytokeratine 5-6 : ref M7237, Dako, Clone II-7, Facteur VIII : ref M0616, Dako, Clone PR36, P53 : ref M7247, Dako, Clone 4A4, Actine : ref M0851, Dako, Clone IA4, Chromogranine : ref A0430, Dako, Clone A, Synaptophysine : ref A0010, Dako, Actine : ref M0851, Dako, 1A4-P53 : ref M0616, Dako, 4A4.</i> </p> <p style="margin-top: 10px;">ADICAP OEGSA7B2, OESGAMB2</p> <div style="text-align: right; margin-top: 20px;"> Dr Alain GAILLOT  </div>			

La Pardieu : Docteurs N. CAUCHOIS-GOUJON, C. DESPLECHAIN, H. EGLOFF, F. FRANCK, M. MOSNIER-DAMET
République : Docteurs A. GAILLOT, F. MAURY, R. VILMANT
Roanne : Secrétariat permanent
Victoria : Docteurs G. LESEC, E. RICHARD-COULET

SELARL CAPITAL 763 000 € - N° ORDRE : 63-25 - E-mail : sipath.pardieu@wanadoo.fr

Page 2 / 2

Figure 71 - Exemple de compte rendu anatomo pathologique - Tumeur mammaire

```
<?xml version="1.0" encoding="iso-8859-1"?>
<EXPORT>18<NURES ID="1250283">
<LIGNE>
<NUDDEEXT>09R020797</NUDDEEXT>
<DATPREL>2009/11/20</DATPREL>
<DATENREG>2009/11/23</DATENREG>
<NUPAT>-1828765</NUPAT>
<NOMPAT>####</NOMPAT>
<PRENOM>####</PRENOM>
<ADRESSE1>####</ADRESSE1>
<ADRESSE2></ADRESSE2>
<ADRESSE3></ADRESSE3>
<CODPOSTAL>####</CODPOSTAL>
<VILLE>####</VILLE>
<CODPAYS></CODPAYS>
<NOMFILLE>####</NOMFILLE>
<SEXE>F</SEXE>
<DATNAISSANCE>####</DATNAISSANCE>
<NOMLEC1>####</NOMLEC1>
<INSEELEC1>631702255</INSEELEC1>
<NOMLEC2></NOMLEC2>
<INSEELEC2>Inconnu</INSEELEC2>
<NOMMED>####</NOMMED>
<INSEEMED></INSEEMED>
<NURES>1250283</NURES>
<DATVALIDATION>2009/11/23</DATVALIDATION>
<MODPREL>B</MODPREL>
<TYPTECH>H</TYPTECH>
<ORGANE>DE</ORGANE>
<LESION>0000</LESION>
<DATCODAGE>2009/11/23</DATCODAGE>
<NOMORIG>POLE SANTE REPUBLIQUE</NOMORIG>
<RESULTATCCL>
</RESULTATCCL>
</LIGNE>
</NURES>
</EXPORT>
```

Figure 72 - Partie non structurée des données informatisées

Mme XX
Née le X

4 RUE XX

XX

Prélevé le 02/11/2009
Reçu le 03/11/2009

Prescrit par le Dr XX

EXAMEN N° 09N219197

Clermont-Ferrand, le 23/11/2009

Dr XX
7 RUE XX

63000 CLERMONT FERRAND

ETUDE COMPLEMENTAIRE PAR HYBRIDATION IN SITU POUR IDENTIFICATION D'UN HPV A RISQUE

RESULTAT : valeur inférieure au seuil de détection

CONCLUSION

Pas d'HPV à risque oncogène identifié par technique d'hybridation in situ.

Procédure Digène hybrid capture II N° série 0922. Test HPV HR NoRTZQ Lot 5272078

Milieu de transport spécifique Digène

Opérateur : FP

Test d'hybridation moléculaire avec amplification du signal chimiohuminescent pour la détection de l'ADN des Papillomas Virus Humains de type oncogène : types 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68.



ADICAP FIGX0I80

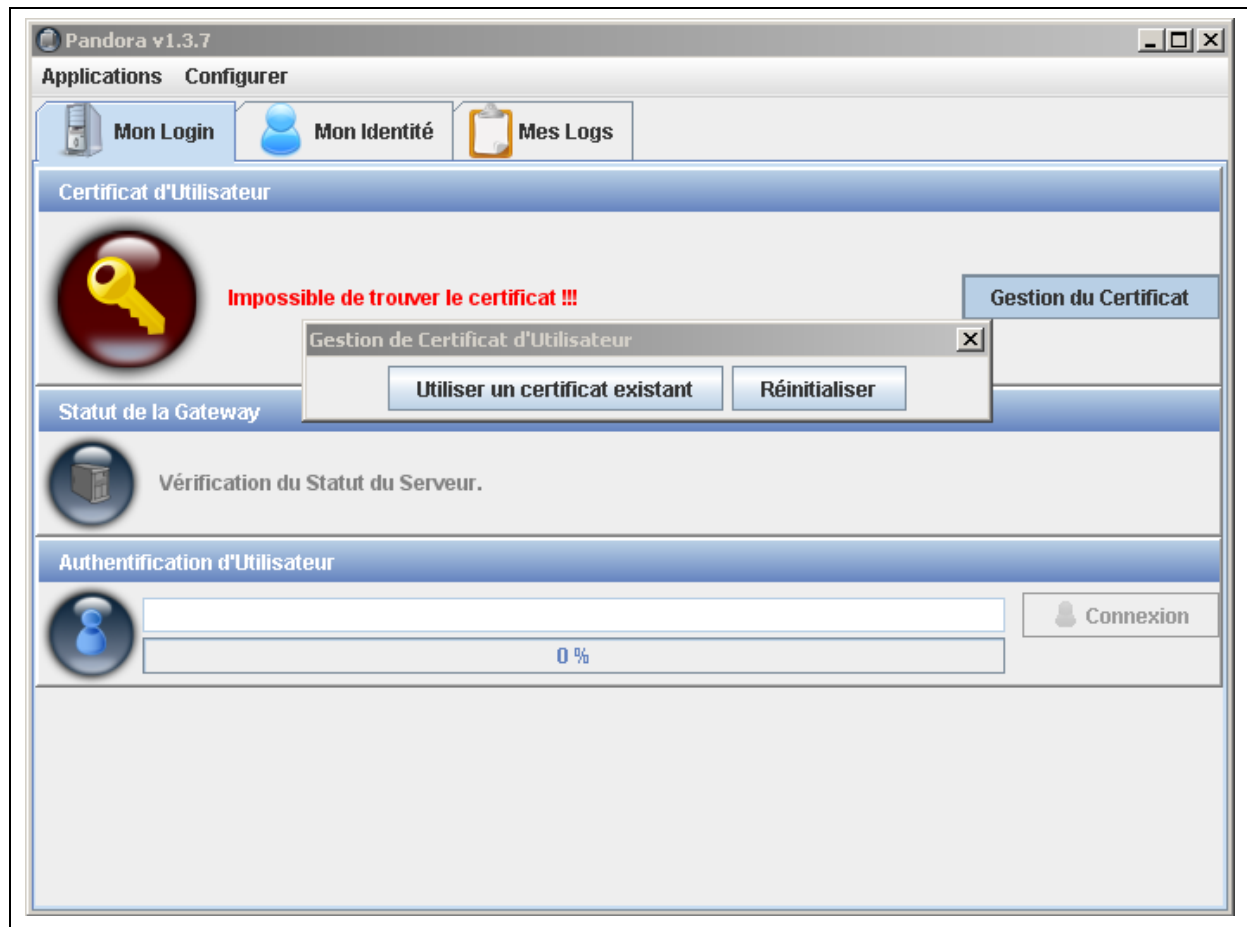
Dr GUY LESEC

.[@NC09N219197]

□

Figure 73 - Partie non structurée des données informatisées

Annexe 2. Interface d'authentification Gateway



Annexe 3. Interface Zeus d'OSI-Santé pour les associations de dépistage

Deux captures d'écran sont fournies du logiciel Zeus, avec l'emplacement du lien qui sera ajouté pour accéder aux données depuis RSCA.

The screenshot shows the Zeus interface for a benign case. The top navigation bar includes 'Fichier', 'Invitations', 'Lecture', 'Suivi', 'Listes', 'Statistiques', 'Compléments', and 'Sécurité'. The main content area is titled 'Fiche anatomo-cytopathologique - Fiche Bénéigne'. It contains a table of examinations with columns 'Date' and 'Examen'. A red arrow points to a button labeled 'Ajout d'un bouton RSCA'.

Figure 74 - Capture d'écran de Zeus, fiche bénigne

The screenshot shows the Zeus interface for a malignant case. The top navigation bar is the same as in Figure 74. The main content area is titled 'Fiche anatomo-cytopathologique - Fiche Maligne - Invasif'. It contains a table of examinations with columns 'Date' and 'Examen'. A red arrow points to a button labeled 'Ajout d'un bouton RSCA'.

Figure 75 - Capture d'écran de Zeus, fiche maligne

Annexe 4. L'algorithme Phonex ^[198]

L'algorithme du Phonex francisé par Brouard est présenté ici en version originale :

- Etape 1 : Remplacer « y » par « i »
- Etape 2 : Supprimer les h s'ils ne sont pas précédés de « s », « c » ou de « p »
- Etape 3 : Remplacer « ph » par « f »
- Etape 4 : Son « an », remplacer :
 - « gan » par « kan »
 - « gam » par « kam »
 - « gain » par « kain »
 - « gaim » par « kaim »
- Etape 5 : Son « yn », remplacer, si suivi par une voyelle
 - « ain », « ein », « aim » et « eim » par « yn »
- Etape 6 : Son « o », remplacer
 - « eau » par « o »
 - « oua » par « 2 »
 - « ein », « ain », « eim » et « aim » par « 4 »
- Etape 7 : Son « é », remplacer
 - « é », « è », « ê », « ai », « ei » par « y »
 - « er » par « yr »
 - « ess » par « yss »
 - « et » par « yt »
- Etape 8 : Sons « an » et « in », remplacer, saufs si suivi d'une voyelle :
 - « an », « am », « en », « em » par « 1 »
 - « in » par « 4 »
- Etape 9 : Son « s », remplacer, s'ils sont encadrés d'une voyelle ou d'un son de 1 à 4 :
- Etape 10 : Remplacer :
 - « s » par « z »
 - « oe », « eu » par « e »
 - « au » par « o »
 - « oi », « oy » par « 2 »
 - « ou » par « 3 »
- Etape 11 : Son « ch », remplacer :
 - « ch », « sch », « sh » par « 5 »
 - « ss », « sc » par « s »
- Etape 12 : remplacer, si suivi d'un « e » ou d'un « i »
 - « c » par « s »
- Etape 13 : remplacer :
 - « c », « q », « qu », « gu » par « k »
 - « ga » par « ka »
 - « go » par « ko »
 - « gy » par « ky »
- Etape 14 : remplacer les lettres suivantes :
 - « a » par « o »
 - « d », « p » par « t »
 - « j » par « g »
 - « b », « v » par « f »
 - « m » par « n »
- Etape 15 : Supprimer les lettres dupliquées
- Etape 16 Supprimer les terminaisons suivantes : t, x
- Etape 17 Affecter à chaque lettres le code numérique correspondant en partant de la dernière lettre :

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	2	3	4	5	e	f	g	h	i	k	l	n	o	r	s	t	u	w	x	y	z

- Etape 18 : Convertir les codes numériques obtenus en un nombre de base 22 exprimé en virgule flottante.

Annexe 5. Comparaison des algorithmes de « data linkage »

Principaux résultats obtenus par Christen lors de la comparaison des algorithmes de « data linkage ». Les chiffres en gras et soulignés indiquent respectivement les meilleurs et moins bons résultats. Les 6 premiers algorithmes sont de type phonétique, les autres sont algébriques.

	Midwives			COMPLETE
	given names	sur- names	full names	surnames
Soundex	.342	.341	.376	.485
Phonex	.423	.369	.499	.579
Phonix	.339	.330	.368	.617
NYSIIS	<u>.275</u>	<u>.296</u>	<u>.299</u>	<u>.351</u>
DMetaphone	.304	.306	.330	.410
FuzSoundex	.327	.311	.359	.396
Leven dist	.658	.513	.737	.624
Dam-L dist	.659	.517	.739	.625
Bag dist	.597	.522	.670	.616
SWater dist	.889	.579	.802	.617
LCS-2	.915	.564	.877	.514
LCS-3	.909	.529	.866	.500
1-grams	.839	.588	.787	.627
2-grams	.885	.498	.867	.519
3-grams	.783	.442	.833	<u>.416</u>
Pos 1-grams	.890	.574	.724	.653
Pos 2-grams	.880	.473	.697	.508
Pos 3-grams	.768	<u>.416</u>	.659	<u>.416</u>
Skip grams	.844	.496	.825	.521
Compr BZ2	<u>.458</u>	.547	<u>.568</u>	.633
Compr ZLib	.532	.456	.684	.481
Jaro	.853	.601	.829	.712
Winkler	.891	.588	.868	.707
SortWink	.803	.580	.809	.707
PermWink	.888	.598	.883	.707
Editex	.631	.561	.706	.646
SAPS dist	.656	.426	.710	.532

Figure 76 - Mesure de probabilité de similarité des différents algorithmes de comparaison de chaînes ;
Crédit Christen [189]

Annexe 6. Version GP-GPU de Jaro-Winkler

L'algorithme de rapprochement d'identités souffre de quelques lenteurs dues notamment au nombre de fois où est appelé la méthode de comparaison de Jaro-Winkler.

Distance de Jaro

L'idée de cet algorithme [192], comme montré en [Figure 77], est de calculer le nombre de caractères communs et des permutations à partir de deux chaînes de caractères fournies en entrée.

Par exemple, pour les deux chaînes « Pierre » et « Peirrick » on obtient :

	P	I	E	R	R	E
P	1	0	0	0	0	0
E	0	0	1	0	0	1
I	0	1	0	0	0	0
R	0	0	0	1	1	0
R	0	0	0	1	1	0
I	0	1	0	0	0	0
C	0	0	0	0	0	0
K	0	0	0	0	0	0

```

Nom: Distance de Jaro
Rôle: Mesure la distance entre deux chaînes de caractères
Entrée: S1 : Chaîne de caractères, S2 : Chaîne de caractères
Sortie: R : Réel
Déclaration: SC1 : Chaîne de caractères
SC2 Chaîne de caractères
t : Entier

début
  SC1 ← CaracteresCommuns(S1,S2)
  SC2 ← CaracteresCommuns(S2,S1)
  t ← 0
  pour i ← 0 à SC1.longueur() faire
    si SC1[i] ≠ SC2[i] alors
      t ← t+0.5
    fin
  finpour
  R ← (SC1.longueur() / S1.longueur() +
    SC2.longueur() / S2.longueur() +
    (SC1.longueur() - t / SC1.longueur()) / 3)
fin
  
```

Figure 77 - Algorithme de calcul de la distance de Jaro

A partir de ce tableau, on détermine les caractères communs aux deux chaînes, ici « PIERR » et « PEIRR » suivant l'ordre des deux chaînes. La longueur C des deux chaînes est alors de 5.

Ensuite, on compte les permutations P qui existent entre ces deux sous-chaînes, au nombre de deux, soit P=2 (une permutation concerne deux caractères). Le calcul de distance peut alors se faire en utilisant la formule :

$$Jaro(PIERRE, PEIRRICK) = \frac{1}{3} \left(\frac{C}{|PIERRE|} + \frac{C}{|PEIRRICK|} + \frac{C-P}{C} \right)$$

$$\text{On obtient, avec : } Jaro(PIERRE, PEIRRICK) = \frac{1}{3} \left(\frac{5}{6} + \frac{5}{8} + \frac{5-2}{5} \right) = 0.752$$

En effectuant une étude sur le fonctionnement de l'algorithme, environ 70% du temps est consacré à la recherche des caractères communs entre les deux chaînes. C'est donc sur ce point que des améliorations sont possibles.

Utilité des processeurs graphiques (GPU)

Depuis la création du GPGPU (General-Purpose Processing on Graphics Processing Units), autrement dit le calcul conventionnel en utilisant les cartes graphiques, les limites du calcul parallèle ont été repoussées. L'architecture des processeurs graphiques, là où un processeur conventionnel contient entre deux et huit unités de calcul, une architecture graphique grand public comme le GF100 de la société Nvidia contient jusqu'à 512 cœurs d'exécution simultanés [209]. Les performances comparées à un CPU sont alors largement supérieures [Figure 78].

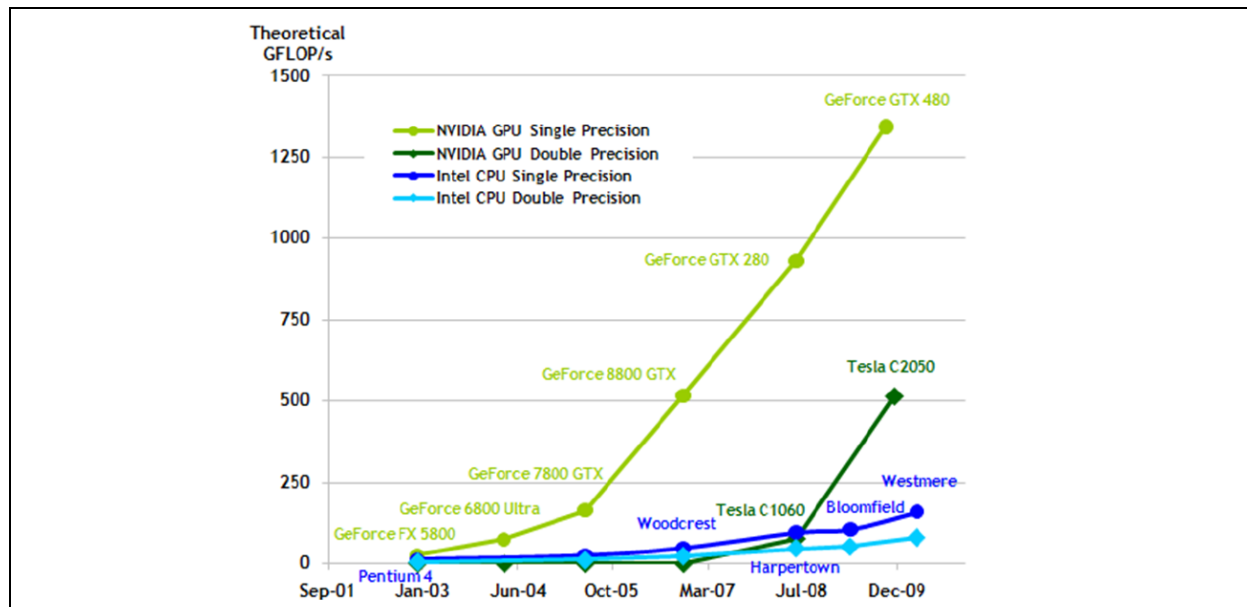


Figure 78 - Comparaison CPU/GPGPU – Crédit Nvidia

Nvidia a aussi proposé un langage de programmation dédié à ces processeurs : CUDA¹. Il s'agit techniquement d'une surcouche de C/C++ fournissant des primitives d'accès aux processeurs graphiques.

Le changement de méthode de programmation est radical comparé à la programmation classique, ou même parallèle. La gestion de la mémoire est très différente car chaque unité de calcul ne dispose que d'une zone de mémoire de travail très limitée. Afin de rendre la programmation parallèle efficace sur un GPGPU, il faut alors « préparer » les données à être traitées de façon très linéaire.

Transposition de Jaro en GPGPU

Dans le cadre du projet RSCA, la taille des chaînes à comparer n'excède que rarement 20 caractères, ce n'est pas sur ce point que le GPGPU sera d'une utilisé mais sur le nombre de comparaisons qui sont à effectuer :

Plutôt que d'effectuer $Jaro(s_1, s_2), Jaro(s_1, s_3) \dots Jaro(s_1, s_n)$, il est possible de calculer toutes les comparaisons s_1, s_n en même temps pour améliorer la vitesse d'exécution globale de Jaro.

Étapes :

- concaténer toutes les chaînes à comparer s_1, s_n dans une chaîne S ;
- créer un tableau de booléens B de taille $|S| * |S|$;
- pour tout couple $(i, j) \in [0, |S|], B[i][j]$, si $S[i] = S[j]$, alors $B[i][j] = vrai$;

Ainsi, le tableau B contient toutes les comparaisons des chaînes en entrée. La recherche de caractères communs dans les chaînes consiste juste à chercher les *vrai* dans le tableau précalculé.

¹ Compute Unified Device Architecture

Tests de performance

L'algorithme présente un avantage de taille : il n'existe aucune relation entre chaque calcul unitaire, ce qui facilite grandement la programmation parallèle. Ainsi, on obtient des résultats assez convaincants sur l'accélération produite par un processeur graphique suivant la taille des données passées à l'algorithme.

En utilisant une configuration de test assez basique, c'est-à-dire une carte graphique mobile de type Nvidia quadro NVS 160M, disposant de 8 cœurs CUDA cadencés à 1450Mhz, comparé à un processeur de type Core 2 Duo P8600 @ 2,4Ghz, le gain de performances varie entre 0.25 et 4 suivant les données en entrée.

En effet, si les données ne sont pas assez volumineuses, les temps d'initialisation sont supérieurs aux temps de calcul. Néanmoins, un facteur 4 est observé lorsque la mémoire est utilisée de façon optimale ce qui laisse de bonnes perspectives.

Bibliographie

- [1] *Dicom - Digital Imaging and Communications in Medicine*. Disponible à: <http://medical.nema.org/>
- [2] *Health Level Seven*. Disponible à: <http://www.hl7.org/>
- [3] *GMSIH - Groupement pour la Modernisation du Systeme d'Information Hospitalier*. Disponible à: <http://www.gmsih.fr/>
- [4] *ATIH - Agence Technique de l'Information sur l'Hospitalisation PMSI - Programme de Médicalisation des Systèmes d'Information*. Disponible à: www.atih.sante.fr
- [5] *ASIP-SANTÉ. Agence des Systèmes d'Information de santé Partagés*. Disponible à: <http://esante.gouv.fr>
- [6] *Décret n° 2010-1229 du 19 octobre 2010 relatif à la télémédecine*. Disponible à: <http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000022932449>
- [7] *Code de la santé publique - Livre III - Chapitre VI*. Disponible à: <http://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000006072665>
- [8] *Loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés*. 2004; Disponible à: <http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006068624>
- [9] *Comité consultatif sur le traitement de l'information en matière de recherche dans le domaine de la santé*. Disponible à: <http://www.enseignementsup-recherche.gouv.fr/cid20537/cctirs.html>
- [10] *CNIL - Recherches biomédicales (Méthodologie de référence n° 1)*. Disponible à: http://www.cnil.fr/fileadmin/documents/declarer/mode_d-emploi/sante/MR-001.pdf
- [11] *Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data*. 1995.
- [12] P. GUARDA. *Data Protection, Information Privacy, and Security Measures: an Essay on the European and the Italian Legal Frameworks*. 2008 Disponible à: <http://eprints.biblio.unitn.it/archive/00001524/>
- [13] J. HERVEG, *Overview on responsibilities in ehealth products and services in European Law*. *Revue du Droit des Technologies de l'Information*, 2007. **29**: p. 273-310.
- [14] *ADICAP - Association pour le Développement de l'Informatique en Cytologie et Anatomopathologie*. Disponible à: <http://www.adicap.asso.fr/>
- [15] WHO. *World Health Organisation - International Classification of Diseases*. Disponible à: <http://www.who.int/classifications/icd>
- [16] INCA (2009) *La situation du cancer en France en 2009*.
- [17] *International Classification of Diseases - ICD10*. 2010 Disponible à: <http://www.who.int/classifications/icd>
- [18] CépiDc - *Centre d'épidémiologie sur les causes médicales de décès*. Disponible à: <http://www.cepiddc.vesinet.inserm.fr/>
- [19] INCA. *Institut National du Cancer*. 2010 Disponible à: <http://www.e-cancer.fr/>
- [20] *Cancer screening in the European Union, Report on the implementation of the Council Recommendation on cancer screening*. 2007 Disponible à:

- http://ec.europa.eu/health/archive/ph_determinants/genetics/documents/cancer_screening.pdf
- [21] Arrêté du 29 septembre 2006 relatif aux programmes de dépistage des cancers 2006; Disponible à: <http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000460656>
- [22] InVs - Taux de participation au programme de dépistage organisé du cancer du sein 2007-2008. 2007 Disponible à: http://www.invs.sante.fr/surveillance/cancers_depistage/participation_depistage_sein_2007_2008.htm
- [23] A. BELOT, ET AL., *Cancer incidence and mortality in France over the period 1980-2005*. Revue d'épidémiologie et de santé publique, 2008. **56**(3): p. 159-175.
- [24] M. COLONNA, ET AL., *Cancer registry data based estimation of regional cancer incidence: application to breast and colorectal cancer in French administrative regions*. British Medical Journal, 1999. **53**(9): p. 558.
- [25] M. ARBYN, P. AUTIER, ET J. FERLAY, *Burden of cervical cancer in the 27 member states of the European Union: estimates for 2004*. Annals of oncology, 2007. **18**(8): p. 1423.
- [26] K. JORGENSEN, P. ZAHL, ET P. GOTZSCHE, *Breast cancer mortality in organised mammography screening in Denmark: comparative study*. British Medical Journal, 2010. **340**.
- [27] P. GØTZSCHE ET M. NIELSEN, *Screening for breast cancer with mammography*. Cochrane Database of Systematic Reviews, 2009. **2**: p. 1-90.
- [28] P. AUTIER ET D. OUAKRIM, *Determinants of the number of mammography units in 31 countries with significant mammography screening*. British journal of cancer, 2008. **99**(7): p. 1185-1190.
- [29] P. AUTIER, ET AL., *Disparities in breast cancer mortality trends between 30 European countries: retrospective trend analysis of WHO mortality database*. British Medical Journal, 2010. **341**: p. c3620.
- [30] C.D. MATHERS, ET AL., *Counting the dead and what they died from: an assessment of the global status of cause of death data*. Bull World Health Organ, 2005. **83**(3): p. 171-7.
- [31] E. CODD, *A relational model of data for large shared data banks*. Communications of the ACM, 1970. **13**(6): p. 377-387.
- [32] W. INMON: *Building the Data Warehouse*. John Wiley & Sons. 1992.
- [33] F. CORBATO ET R. FANO, *Time-sharing on Computers*. Information, a Scientific American Book, 1966.
- [34] L. KLEINROCK, *UCLA to be first station in nationwide computer network*. UCLA, Office of Public Information, 1969. **3**.
- [35] G. MOORE, *Cramming more components onto integrated circuits*. Electronics, 1965. **38**(8).
- [36] P. KOCOVIC, *Four laws for today and tomorrow*. Journal of Applied Research and Technology, 2009. **6**(03).
- [37] I. FOSTER ET C. KESSELMAN, *The Grid: Blueprint for a New Computing Infrastructure*. 1999. 1999, Morgan Kaufmann.
- [38] I. FOSTER, C. KESSELMAN, ET S. TUECKE, *The anatomy of the grid: Enabling scalable virtual organizations*. International Journal of High Performance Computing Applications, 2001. **15**(3): p. 200.
- [39] I. FOSTER, *What is the grid? a three point checklist*. GRID today, 2002. **1**(6): p. 32-36.
- [40] H. STOCKINGER, *Defining the grid: a snapshot on the current view*. The Journal of Supercomputing, 2007. **42**(1): p. 3-17.
- [41] CERN: *Organisation Européenne pour la Recherche Nucléaire*. Disponible à: www.cern.ch
- [42] LHC: *Large Hadron Collider*. Disponible à: <http://lhc.web.cern.ch>
- [43] LCG: *LHC Computing Grid*. Disponible à: <http://lhc.web.cern.ch>
- [44] *The Datagrid project*. Disponible à: <http://eu-datagrid.web.cern.ch/eu-datagrid/>
- [45] W. HOSCHEK, ET AL., *Data management in an international data grid project*. Journal of Grid Computing, 2000: p. 333-361.
- [46] EGEE - *Enabling Grids for E-Science*. Disponible à: <http://www.eu-egee.org/>

- [47] *European Grid Initiative*. 2010 Disponible à: <http://www.egi.eu>
- [48] *Réseau GÉANT*. Disponible à: <http://www.geant.net>
- [49] *GIP Renater*. Disponible à: <http://www.renater.fr/>
- [50] J. GEELAN, *The Top 150 Players in Cloud Computing*, in *Virtualisation Conference*. 2010: New York.
- [51] R. BOLZE, ET AL., *Grid'5000: a large scale and highly reconfigurable experimental grid testbed*. International Journal of High Performance Computing Applications, 2006. **20**(4): p. 481.
- [52] *INRIA - Inédit : De grid 5000 à Aladdin*. Disponible à: <http://www.pyoudeyer.com/ineditINRIAOUdeyerFrench08.pdf>
- [53] S. MATSUOKA, ET AL., *Japanese computational grid research project: NAREGI*. Proceedings of the IEEE, 2005. **93**(3): p. 522-533.
- [54] A. CHERVENAK, ET AL., *The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets*. Journal of Network and Computer Applications, 2000. **23**(3): p. 187-200.
- [55] *Open Science Grid*. Disponible à: <http://www.opensciencegrid.org/>
- [56] R. PORDES, ET AL. *The open science grid*, 2007: IOP Publishing.
- [57] P. BECKMAN, *Building the TeraGrid*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2005. **363**(1833): p. 1715.
- [58] *Teragrid 10*. 2010 Disponible à: <https://www.teragrid.org/web/events/tg10/presentations>
- [59] P. AVERY, *Open Science Grid: Linking Universities and Laboratories in National Cyberinfrastructure*, in *Teragrid*. 2009: Arlington.
- [60] H. LEDERER ET V. ALESSANDRINI, *DEISA: Enabling Cooperative Extreme Computing in Europe*. Parallel Computing: Architectures, Algorithms and Applications, 2008. **15**: p. 978-1.
- [61] *Top 500 Supercomputing sites*. 2010 Disponible à: <http://www.top500.org/list/2010/06/100>
- [62] *Gnutella* Disponible à: <http://rfc-gnutella.sourceforge.net/>
- [63] J. LIANG, R. KUMAR, ET K. ROSS, *Understanding kazaa*. Manuscript, Polytechnic Univ, 2004.
- [64] O. HECKMANN ET A. BOCK, *The edonkey 2000 protocol*. 2002, Technical Report KOM-TR-08-2002, Multimedia Communications Lab, Darmstadt University of Technology.
- [65] *Bittorrent*. Disponible à: <http://www.bittorrent.com/>
- [66] D. ANDERSON, ET AL., *SETI@ home: an experiment in public-resource computing*. Communications of the ACM, 2002. **45**(11): p. 56-61.
- [67] E.S. REICH, *Mysterious signals from light years away* in *newscientist.com*. 2004.
- [68] *Boinc*. Disponible à: <http://boinc.berkeley.edu/>
- [69] *Boinc Stats, Octobre 2010*. Disponible à: http://boincstats.com/stats/project_graph.php
- [70] G. FEDAK, ET AL. *Xtremweb: A generic global computing system*. in *IEEE/ACM International Symposium on Cluster Computing and the Grid*, Brisbane, 2001.
- [71] A. BEBERG, ET AL. *Folding@ home: Lessons from eight years of volunteer distributed computing*. in *IEEE International Symposium on Parallel&Distributed Processing*, Rome, 2009.
- [72] E. URBAN, ET AL., *Edges: Bridging egee to boinc and xtremweb*. Journal of Grid Computing, 2009. **7**(3): p. 335-354.
- [73] R. WARREN, ET AL., *MammoGrid: A prototype distributed mammographic database for Europe*. Clinical radiology, 2007. **62**(11): p. 1044-1051.
- [74] J. GRETHE, ET AL., *Biomedical informatics research network: building a national collaboratory to hasten the derivation of new understanding and treatment of disease*. Studies in health technology and informatics, 2005. **112**: p. 100-110.
- [75] N. WALTON, ET AL., *AstroGrid: A place for your science*. Astronomy & Geophysics, 2006. **47**(3): p. 3.22.
- [76] *Gridchem - Computational Chemistry Grid*. Disponible à: <https://www.gridchem.org/>
- [77] *Atlas Experiment - Cern*. Disponible à: <http://atlas.web.cern.ch>
- [78] *D-Grid - Deutsche Grid Initiative*. Disponible à: <http://www.d-grid.de/>
- [79] *Auvergrid*. Disponible à: <http://www.auvergrid.fr/>
- [80] *Scotgrid*. Disponible à: <http://www.scotgrid.ac.uk/>
- [81] *Ibergrid*. Disponible à: <http://www.ibergrid.eu/>

- [82] *EUAsiaGrid*. Disponible à: <http://www.euasiagrid.org/>
- [83] J. HERVEG, *Does HealthGrid Present Specific Risks With Regard To Data Protection? From genes to personalized healthcare: grid solutions for the life sciences*, 2007: p. 219.
- [84] *Lifegrid*. Disponible à: www.lifegrid.fr
- [85] M. DIARENA, ET AL., *HOPE, an open platform for medical data management on the grid*. Stud Health Technol Inform, 2008. **138**: p. 34-48.
- [86] *Edital - Echange de Données Informatiques de Traçabilité en région Auvergne-Loire*. Disponible à: <http://www.edital.fr/>
- [87] J. MONTAGNAT, ET AL., *NeuroLOG: a community-driven middleware design*. Stud Health Technol Inform, 2008: p. 49-58.
- [88] *Inria - Rennes - Specification of the NeuroLOG architecture components - Deliverable L3*. 2007 Disponible à: <http://neurolog.polytech.unice.fr>
- [89] J. FREUND, ET AL., *Health-e-child: an integrated biomedical platform for grid-based paediatric applications*. Stud Health Technol Inform, 2006. **120**: p. 259-70.
- [90] M. BRADY, ET AL., *eDiamond: a Grid-enabled federated database of annotated mammograms*. Grid Computing: Making the Global Infrastructure a Reality, 2003. **39**.
- [91] V. BRETON, K. DEAN, ET T. SOLOMONIDES, *The Healthgrid white paper*. Studies in health technology and informatics, 2005. **112**: p. 249.
- [92] *Healthgrid association*. Disponible à: www.healthgrid.org
- [93] M. TSIGNAKIS, ET AL., *Building a European biomedical grid on cancer: the ACGT Integrated Project*. Stud Health Technol Inform, 2006. **120**: p. 247-58.
- [94] R. HAUX, ET AL.: *Strategic information management in hospitals: an introduction to hospital information systems*. 2004: Springer Verlag.
- [95] ISO. *Technical Committee ISO/TC 215 HI. ISO/TR 20514 Health Informatics - Electronic Health Record - Definition, scope and context*. 2005;
- [96] AFNOR. *Commission de Normalisation de l'Informatique de la Santé et de l'Action Sociale : AFNOR/S95N*. 2008; Disponible à: http://www2.afnor.org/espace_normalisation/structure.aspx?commid=1816
- [97] ARS. *Lancement du DMP en fin d'année annoncé par le Ministre de la Santé*. 2010 Disponible à: <http://ars.sante.fr/Roselyne-Bachelot-confirme-le.96162.0.html>
- [98] *Système MultiSource Cancer*. Disponible à: <http://www.invs.sante.fr/surveillance/cancers/travaux.htm>
- [99] CNIL - *Etat des lieux en matière de procédés d'anonymisation*. Disponible à: <http://www.cnil.fr/en-savoir-plus/fiches-pratiques/fiche/article/letat-des-lieux-en-matiere-de-procedes-danonymisation>
- [100] ASIP-SANTÉ. *L'interopérabilité des données de santé : Comment passer à l'acte ?* in *Journées Francophones d'Informatique Médicale (JFIM)*, Nice, 2009.
- [101] ASIP-SANTÉ. *Cahier des charges de la DMP-Compatibilité*. Disponible à: <http://esante.gouv.fr/contenu/asip-sante-publication-du-cahier-des-charges-de-la-dmp-compatibilite-actualise>
- [102] ASIP-SANTÉ. *Hébergeurs agréés*. Disponible à: <http://esante.gouv.fr/contenu/hebergeurs-agrees>
- [103] *Dossier Communicant de Santé et Dossier Médical Partagé*. Disponible à: http://esante.gouv.fr/sites/default/files/101015_PhSimian_Ppt_DCC.pdf
- [104] *Plan Cancer 2009-2013*. Disponible à: http://www.e-cancer.fr/component/docman/doc_download/3855-brochure-plan-cancer-2009-2013+plan+cancer
- [105] OPNEHR, *Archetypes definition and principles*.
- [106] HL7 - *Reference Information Model (RIM)*. Disponible à: <http://www.hl7.org/implement/standards/rim.cfm>
- [107] CEN - *Comité Européen de Normalisation*. Disponible à: <http://www.cen.eu>
- [108] CEN - *Technical Committee 251: Health Informatics*. Disponible à: <http://www.centc251.org>
- [109] CEN/ISO 13606 *Information*. Disponible à: <http://www.en13606.org/>

- [110] P. SCHLOEFFEL, ET AL. *The relationship between CEN 13606, HL7, and openEHR*. in *HIC and HINZ*, 2006.
- [111] *Gstat - EGEE sites Summary*. Disponible à: <http://gstat-prod.cern.ch/gstat/summary/GRID/EGEE/>
- [112] *Groupeement d'Interêt Scientifique - France Grilles*.
- [113] *neuGRID*. Disponible à: www.neugrid.eu
- [114] *e-NMR*. Disponible à: www.enmr.eu
- [115] L. MAIGNE: *Personalized dosimetry using GATE Monte Carlo simulations on grid architecture : Application in ocular brachytherapy*. Thèse de doctorat, Université Blaise Pascal, 2005
- [116] C. TIAM: *Dosimétrie en radiothérapie et curiethérapie par simulation Monte-Carlo GATE sur grille informatique*. Thèse de doctorat, Université Blaise Pascal, 2007
- [117] D. LINGRAND, T. GLATARD, ET J. MONTAGNAT, *Modeling the latency on production grids with respect to the execution context*. *Parallel Computing*, 2009. **35**(10-11): p. 493-511.
- [118] *gLite - Lightweight middleware for grid computing*. Disponible à: <http://glite.web.cern.ch/glite/>
- [119] *The Globus Alliance*. Disponible à: <http://www.globus.org/>
- [120] C. ADAMS ET S. LLOYD: *Understanding PKI: concepts, standards, and deployment considerations*. 2002: Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.
- [121] RSA-LABORATORIES, *RFC 3280 - Internet X.509 Public Key Infrastructure - Certificate and Certificate Revocation List (CRL) Profile*. 2002.
- [122] F. VON WELCH, ET AL., *Security for grid services*. Arxiv preprint cs/0306129, 2003.
- [123] P. ZIMMERMANN: *The official PGP user's guide*. 1995.
- [124] *gLite User guide - Version 3.1*. Disponible à: <https://edms.cern.ch/file/722398/1.3/gLite-3-UserGuide.pdf>
- [125] *Berkeley Database Information Index V5*. Disponible à: <https://twiki.cern.ch/twiki/bin/view/EGEE/BDII>
- [126] J. BROOKE, ET AL. *Semantic matching of grid resource descriptions*. in *2nd European Across-Grids Conference*, Chypre, 2004.
- [127] P. ANDREETTO, ET AL., *The gLite workload management system*. *Journal of Physics, Conference Series* 119, 2008. **119**(6).
- [128] A. ANJOMSHOAA, ET AL. *Job submission description language (jsdl) specification, version 1.0*. in *Open Grid Forum*, 2005.
- [129] W. ALLCOCK, ET AL. *The Globus striped GridFTP framework and server*. in *SC'05, Seattle, Seattle*, 2005.
- [130] J. BAUD, ET AL. *Lcg data management: From edg to egee*. in *eScience All Hands Meeting*, Nottingham, 2005.
- [131] C. MUNRO ET B. KOBLITZ, *Performance comparison of the LCG2 and gLite file catalogues*. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 2006. **559**(1): p. 48-52.
- [132] *EGEE JRA1 - Fireman Catalog User Guide*, J.D.M. Cluster, Editor. 2005.
- [133] B. KOBLITZ, N. SANTOS, ET V. POSE, *The amga metadata service*. *Journal of Grid Computing*, 2008. **6**(1): p. 61-76.
- [134] V. KASAM, ET AL., *WISDOM-II: Screening against multiple targets implicated in malaria using computational grid infrastructures*. *Malaria Journal*, 2009. **8**(1): p. 88.
- [135] J. MONTAGNAT, ET AL., *Bridging clinical information systems and grid middleware: a Medical Data Manager*. *Studies in health technology and informatics*, 2006. **120**: p. 14.
- [136] M. GUDGIN, ET AL., *SOAP Version 1.2*. W3C Working Draft, 2001. **9**.
- [137] E. CHRISTENSEN, ET AL., *Web services description language (WSDL) 1.1*. 2001, W3C.
- [138] L. CLEMENT, ET AL., *UDDI Version 3.0. 2*. UDDI Spec Technical Committee Draft, 2004. **10**.
- [139] F. SULLIVAN. *Smart Cards for Healthcare in Europe*. 2010 Disponible à: <http://www.frost.com/prod/servlet/market-insight-print.pag?docid=200942088>
- [140] GIP-CPS. *Guide de présentation CPSv3 et des services associés*. Disponible à: https://editeurs.gip-cps.fr/documents/techniques/GIP-CPS_Presentation-CPS3_v1.2.pdf

- [141] GIP-CPS. *Les certificats X.509 et les CRLs des cartes CPS*. Disponible à: https://editeurs.gip-cps.fr/documents/techniques/CPS2ter_Certificats_X-509_V1-4c.pdf
- [142] GIP-CPS, *Installation CryptolibCPS Windows*.
- [143] GIP-CPS. *Cryptolib-CPS sur lecteurs PC/SC*. Disponible à: <https://editeurs.gip-cps.fr/documents/techniques/Installation-Cryptolib-Win32.pdf>
- [144] GIP-CPS. *Annuaire des Autorités de Certification et CRL*. Disponible à: <http://annuaire.gip-cps.fr/>
- [145] RSA Laboratories - Public-Key Cryptography Standards (PKCS). Disponible à: <http://www.rsa.com/rsalabs/node.asp?id=2124>
- [146] maat-Gknowledge. Disponible à: <http://www.g-knowledge.com/>
- [147] Health-e-Child awards. Disponible à: <http://www.health-e-child.org/awards>
- [148] Java Grid Application Toolkit (GAT). Disponible à: <http://www.cs.vu.nl/ibis/javagat.html>
- [149] S. LE BLOND, A. OPRESCU, ET C. ZHANG. *Early application experience with the grid application toolkit (gat)*. in *Workshop on Grid Applications held in conjunction with the Fourteenth Global Grid Forum (GGF'14)*, Chicago, 2005.
- [150] *Simple Api for Grid Applications (SAGA)*. Disponible à: <http://saga.cct.lsu.edu/>
- [151] S. JHA, ET AL. *Grid Interoperability at the application level using SAGA*. in *IEEE International Conference on e-Science and Grid Computing*, Bangalore, 2008.
- [152] J. ROSENBERG ET D. REMY: *Securing Web Services with WS-Security: Demystifying WS-Security, WS-Policy, SAML, XML Signature, and XML Encryption*. 2004: Pearson Higher Education.
- [153] J. EPSTEIN, S. MATSUMOTO, ET G. MCGRAW, *Software security and SOA: danger, Will Robinson!* Security & Privacy, IEEE, 2006. **4**(1): p. 80-83.
- [154] G. ALOISIO, ET AL., *Secure web services with Globus GSI and gSOAP*. Euro-Par 2003 Parallel Processing, 2004: p. 421-426.
- [155] *The Globus Toolkit 4 Programmer's Tutorial*. Disponible à: <http://gdp.globus.org/gt4-tutorial/>
- [156] UMLS - Unified Medical Language System. Disponible à: <http://www.nlm.nih.gov/research/umls/>
- [157] Galen. Disponible à: www.galen.org
- [158] Gene Ontology. Disponible à: www.geneontology.org
- [159] Infologic-Santé. Disponible à: <http://www.infologic-sante.fr>
- [160] Diamic. Disponible à: <http://www.infologic-sante.fr/diamic-presentation.php>
- [161] Haute Autorité de Santé. Disponible à: <http://www.has-sante.fr>
- [162] P. FORTUIT, *Professional health cards (CPS): informatic health care system in France*. Ann Pharm Fr, 2005. **63**(5): p. 350-5.
- [163] ASIP-SANTÉ. *Identifiant National de Santé; spécification 07/2010*. Disponible à: <http://esante.gouv.fr/referentiels/identification/dossier-de-conception-de-l-identifiant-national-de-sante-calculé-ins-c>
- [164] CNIL. *L'usage du NIR pour l'identification des données*. Disponible à: <http://www.cnil.fr/en/la-cnil/actu-cnil/article/article/le-nir-un-numero-dont-lusage-doit-rester-cantonne/>
- [165] A. SYALIM, Y. HORI, ET K. SAKURAI. *Comparison of Risk Analysis Methods: Mehari, Magerit, NIST800-30 and Microsoft's Security Management Guide*. in *International Conference on Availability, Reliability and Security*, 2009.
- [166] CLUSIF. *Guide de l'analyse des risques - Mehari*. Disponible à: http://www.clusif.asso.fr/fr/production/ouvrages/pdf/Guide_AnalyseRisques.pdf
- [167] J. PASSERAT-PALMBACH, *Mise en place d'un environnement de sécurité autour de la Carte de Professionnel de Santé dans une infrastructure de grille de données*. 2009, Rapport d'ingénieur, Filière : Informatique des Systèmes Embarqués, Université Blaise Pascal - Clermont-Ferrand.
- [168] RSA Laboratories. Disponible à: <http://www.rsa.com/rsalabs/>
- [169] PKCS#11 Standard - RSA Laboratories. Disponible à: <http://www.rsa.com/rsalabs/node.asp?id=2133>
- [170] Mozilla: Network Security Services (NSS). Disponible à: <http://www.mozilla.org/projects/security/pki/nss/>

- [171] *Décret n° 46-1432 du 14 juin 1946 [...] relatif à l'INSEE*. 1946;
- [172] C. QUANTIN ET K. BOURQUARD, *Identification du patient dans les systèmes d'information de santé*. École d'Été Méditerranéenne d'Information en Santé, Corte, 2007.
- [173] *Health and Human Services Department - Health Information Privacy*. Disponible à : <http://www.hhs.gov/ocr/privacy/>
- [174] P. THOMAS ET C. EVANS, *An identity crisis? Aspects of patient misidentification*. *Clinical Risk*, 2004. **10**(1): p. 18.
- [175] ANSSI: *Agence Nationale de la Sécurité des Systèmes d'Information*. Disponible à : <http://www.ssi.gouv.fr/>
- [176] ASIP-SANTÉ. *Algorithme de calcul de l'INS-C*. 2009 Disponible à : http://www.i-med.fr/IMG/pdf/Dossier_de_conception_INS-C_-_Algorithme_de_calcul_v0.0.1.pdf
- [177] P. LEACH, M. MEALLING, ET R. SALZ, *A Universally Unique Identifier (UUID) URN Namespace*. 2005, RFC 4122.
- [178] C. FRIEDMAN ET R. SIDELI, *Tolerating spelling errors during patient validation*. *Computers and Biomedical Research*, 1992. **25**(5): p. 486-509.
- [179] H.L. DUNN, *Record linkage*. *American Journal of Public Health*, 1946. **36**(12): p. 1412.
- [180] H.B. NEWCOMBE, *Record linking: the design of efficient systems for linking records into individual and family histories*. *American Journal of Human Genetics*, 1967. **19**(3 Pt 1): p. 335.
- [181] E.D. ACHESON, *Medical record linkage*. *Methods of information in medicine*, 1969. **8**(1): p. 1.
- [182] F.J. DAMERAU, *A technique for computer detection and correction of spelling errors*. *Communications of the ACM*, 1964. **7**(3): p. 171-176.
- [183] J.L. PETERSON, *A note on undetected typing errors*. *Communications of the ACM*, 1986. **29**(7): p. 633-637.
- [184] K. KUKICH, *Techniques for automatically correcting words in text*. *ACM Computing Surveys (CSUR)*, 1992. **24**(4): p. 439.
- [185] J.J. POLLOCK ET A. ZAMORA, *Automatic spelling correction in scientific and scholarly text*. *Communications of the ACM*, 1984. **27**(4): p. 358-368.
- [186] W.E. WINKLER, *The state of record linkage and current research problems*. Technical Report, Statistical Research Division, U.S. Bureau of the Census, 1999.
- [187] M. DU, *Approximate name matching*. Master's thesis, Royal Institute of Technology, Stockholm, 2005.
- [188] W.E. WINKLER, *Overview of record linkage and current research directions*. Research Report Series, 2006. **2**.
- [189] P. CHRISTEN, *A comparison of personal name matching: Techniques and practical issues*. Data Mining Workshops, International Conference on, 2006: p. 290-294.
- [190] W. COHEN, P. RAVIKUMAR, ET S. FIENBERG. *A comparison of string distance metrics for name-matching tasks*. in *IJCAI Workshop on Information Integration on the Web*, Acapulco, 2003.
- [191] R.W. HAMMING, *Error detecting and error correcting codes*. *Bell System Technical Journal*, 1950. **29**(2): p. 147-160.
- [192] M.A. JARO, *Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida*. *Journal of the American Statistical Association*, 1989. **84**(406): p. 414-420.
- [193] R. RUSSELL, *United States patent 1435663*. Washington, DC: United States Patent Office, 1922.
- [194] D.E. KNUTH: *The Art of Computer Programming 1: Fundamental Algorithms 2: Seminumerical Algorithms 3: Sorting and Searching*. 1968: Addison-Wesley.
- [195] J.C. HERMANSEN, D. LORITZ, ET R. O'BRIEN: *Automatic name searching in large data bases of international names*. 1985: Georgetown University.
- [196] F. PATMAN ET L. SHAEFER, *Is Soundex good enough for you? On the hidden risks of Soundex-based name searching*. Language Analysis Systems, Inc., Herndon, 2001.
- [197] A. LAIT ET B. RANDELL, *An assessment of name matching algorithms*. Technical Report Series - University of Newcastle Upon Tyne, 1996.
- [198] F. BROUARD. *L'algorithme "Phonex" francisé*. 1999 Disponible à : <http://sqlpro.developpez.com/cours/soundex/>

- [199] S. CHAPMAN. *SimMetrics: a Java / C# / .NET library of Similarity Metrics*. 2005 Disponible à : <http://sourceforge.net/projects/simmetrics/>
- [200] M.A. JARO, *Probabilistic linkage of large public health data files*. Statistics in medicine, 1995. **14**(5 7): p. 491-498.
- [201] S. KENDRICK ET J. CLARKE, *The Scottish Record Linkage System*. Health Bulletin, 1993. **51**(2): p. 72.
- [202] L.E. GILL. *OX-LINK: The Oxford Medical Record Linkage System*, 1997: National Academies.
- [203] C.D.A.J. HOLMAN, ET AL., *Population based linkage of health records in Western Australia: development of a health services research linked database*. Australian and New Zealand Journal of Public Health, 1999. **23**(5): p. 453-459.
- [204] P. CHRISTEN. *Febrl: an open source data cleaning, deduplication and record linkage system with a graphical user interface*. in *KDD '08*, New York, 2008: ACM.
- [205] E. SAULEAU, J. PAUMIER, ET A. BUEMI, *Medical record linkage in health information systems by approximate string matching and clustering*. BMC Medical Informatics and Decision Making, 2005. **5**(1): p. 32.
- [206] J. PAUMIER, E. SAULEAU, ET A. BUEMI. *Partage de données médicales: Agrégats d'identités approchantes*. in *Journées Francophones d'Informatique Médicale*, Lille, 2005.
- [207] E. GAMMA, ET AL.: *Design patterns: elements of reusable object-oriented software*. Vol. 206. 1995: Addison-wesley Reading, MA.
- [208] *Social Security Death Index*. Disponible à : <http://ssdi.rootsweb.ancestry.com>
- [209] *Nvidia GF100 whitepaper*. Disponible à : http://www.nvidia.fr/object/IO_86775.html

Liste des publications

- [210] M. DIARENA, S. NOWAK, J.Y. BOIRE, V. BLOCH, D. DONNARIEIX, A. FESSY, B. GRENIER, B. IRRTHUM, Y. LEGRE, L. MAIGNE, J. SALZEMANN, C. THIAM, N. SPALINGER, N. VERHAEGHE, P. DE VIEGER, ET V. BRETON, *HOPE, an open platform for medical data management on the grid*. Stud Health Technol Inform, 2008. **138**: p. 34-48.
- [211] C. QUANTIN, G. COATRIEUX, M. FASSA, V. BRETON, D.O. JAQUET-CHIFFELLE, P. DE VIEGER, N. LYPZYC, J.Y. BOIRE, C. ROUX, ET F.A. ALLAERT, *Centralised versus decentralised management of patients' medical records*. Stud Health Technol Inform, 2009. **150**: p. 700-4.
- [212] C. QUANTIN, G. COATRIEUX, F.A. ALLAERT, M. FASSA, K. BOURQUARD, J.Y. BOIRE, P. DE VIEGER, L. MAIGNE, ET V. BRETON, *New advanced technologies to provide decentralised and secure access to medical records: case studies in oncology*. Cancer Inform, 2009. **7**: p. 217-29.
- [213] P. DE VIEGER, J.Y. BOIRE, V. BRETON, Y. LEGRE, D. MANSSET, J. REVILLARD, D. SARRAMIA, ET L. MAIGNE, *Grid-enabled sentinel network for cancer surveillance*. Stud Health Technol Inform, 2009. **147**: p. 289-94.
- [214] V. BRETON, A.L. DA COSTA, P. DE VIEGER, Y.M. KIM, L. MAIGNE, R. REUILLON, D. SARRAMIA, N.H. TRUONG, H.Q. NGUYEN, D. KIM, ET Y.T. WU, *Innovative in silico approaches to address avian flu using grid technology*. Infect Disord Drug Targets, 2009. **9**(3): p. 358-65.
- [215] P. DE VIEGER, J.Y. BOIRE, V. BRETON, Y. LEGRE, D. MANSSET, J. REVILLARD, D. SARRAMIA, ET L. MAIGNE, *Sentinel e-health network on grid: developments and challenges*. Stud Health Technol Inform, 2010. **159**: p. 134-45.
- [216] P. DE VIEGER, J.Y. BOIRE, V. BRETON, G.M. A, L. MAIGNE, ET D. MANSSET, *Cancer Sentinel project: a regional network to enable Cancer data exchange in Auvergne*. Bulletin du Cancer, 2010. **97**: p. 49-50.

Table des figures

Figure 1 - Organisation Européenne de la protection des données	23
Figure 2 - Les registres des cancers en France. Source Francim	28
Figure 3 - Cas de cancer du sein, colon et col utérin en UE (2007) [20].....	30
Figure 4 - Décès dus aux cancers du sein, colon et col en UE (2007) [20]	30
Figure 5 - Estimation de la population française ciblée par le dépistage organisé des cancers	31
Figure 6 - Diagramme de séquence du dépistage organisé du cancer du sein	32
Figure 7 - Comparaison des lois de Moore, Gilder et Metcalfe	35
Figure 8 - Architecture d'une grille	37
Figure 9 - Topologie GÉANT en Europe - Source GÉANT2.net	38
Figure 10 - Réseau GÉANT et débits théoriques mondiaux - Source GÉANT2.net.....	39
Figure 11 - Réseau Renater - Source Renater.fr.....	39
Figure 12 - Interactions CE-SE lors de l'exécution d'un job.....	41
Figure 13 - Représentation schématique des organisations virtuelles	45
Figure 14 - Système déclaratif / système connecté	48
Figure 15 - Système de collecte	96
Figure 16 - Séquence d'envoi déclaratif de données	96
Figure 17 - Exemple de diagramme UML - Gestion des rôles par HL7-RIM - Source HL7.org.....	100
Figure 18 - Document minimal HL7-CDA	100
Figure 19 - Relations openEHR/CEN/HL7	101
Figure 20 - Chronologie DataGrid - EGEE - EGI.....	102
Figure 21 - Organisation des tâches EGI-NGI	103
Figure 22 - Evolution du temps de calcul d'une simulation par rapport au nombre de jobs sur grille - source [116]	104
Figure 23 - Composants gLite.....	106
Figure 24 - Exemple de certificat – Vue Kleopatra	107
Figure 25 - Diagramme de séquence : Etablissement d'une connexion par certificat	108
Figure 26 - RFC3280 : Motifs de révocation des certificat	108
Figure 27 - Architecture de VOMS	109
Figure 28 - Interface d'administration de VOMS	109
Figure 29 - Composition des groupes dans la VO Sentinelle.....	110
Figure 30 - Exemple de fichier JDL	112
Figure 31 - Cycle de vie d'un Job	113
Figure 32 - Fonctionnement du LFC pour le stockage d'un fichier	114
Figure 33 - Constitution de la partie administration de la VO Sentinelle	117
Figure 34- Contenu d'une carte CPS - Source Gip-CPS.....	121
Figure 35 - Couches PKCS#11 pour CPS	122
Figure 36 - Architecture globale de Pandora Gateway - Crédit maat-G	124

Figure 37 - Sécurité des services Gateway - Crédit maat-G	125
Figure 38 - Diagramme de séquence - Service d'authentification Gateway	126
Figure 39 - Système « <i>Integrated Case Data</i> »	127
Figure 40 - Méthodologie d'import des données dans ICD.....	128
Figure 41 - Architecture complète du réseau	129
Figure 42 - Système de réplication et de collaboration dans AMGA	130
Figure 43 - Modèle de données - Compte rendu anatomo-pathologique.....	132
Figure 44 - Interfaçage avec les Logiciels Existants.....	134
Figure 45 - Interfaçage avec les logiciels métier – Point de vue dépistage.....	135
Figure 46 - Interfaçage du réseau sentinelle avec la Santé publique.....	135
Figure 47 - Fonctionnement global de Mehari – Crédit : Clusif	142
Figure 48 - Récapitulatif de l'analyse des risques	143
Figure 49 - Architecture des ACs du GIP-CPS	145
Figure 50 - Versions de PKCS.....	146
Figure 51 - Scénario d'authentification CPS	147
Figure 52 - Schéma en couches de la communication avec la CPS	147
Figure 53 - Identifiants et versions chiffrées.....	153
Figure 54 - Problème d'identification	153
Figure 55 - Arbre d'identification pour un patient.....	154
Figure 56 - Fonctionnement du système d'identification	158
Figure 57 - Algorithme de calcul de la distance de Levensthein	162
Figure 58 - Algorithme de calcul de la distance de Jaro.....	163
Figure 59 - Table de correspondance lettres-code soundex en anglais.....	164
Figure 60 - Soundex en français	164
Figure 61 - Problème théorique de rapprochement d'identités.....	166
Figure 62 - Problème pratique de rapprochement d'identités	166
Figure 63 - Processus d'identification	167
Figure 64 - Diagramme de classes - Compateur de champs	168
Figure 65 - Paramètres utilisés pour la comparaison.....	170
Figure 66 - Résultats bruts de data linkage.....	170
Figure 67 - Résultats combinés de data linkage.....	171
Figure 68 - Résultat de comparaison en fonction du seuil.....	172
Figure 69 - Résultats de comparaison - Date de naissance et adresse incluse	173
Figure 70 - Algorithme simplifié de matching	174
Figure 71 - Exemple de compte rendu anatomo pathologique - Tumeur mammaire	182
Figure 72 - Partie non structurée des données informatisées	183
Figure 73 - Partie non structurée des données informatisées	184
Figure 74 - Capture d'écran de Zeus, fiche bénigne	186
Figure 75 - Capture d'écran de Zeus, fiche maligne	186
Figure 76 - Mesure de probabilité de similarité des différents algorithmes de comparaison de chaînes ; Crédit Christen [189].....	188
Figure 77 - Algorithme de calcul de la distance de Jaro.....	189
Figure 78 - Comparaison CPU/GPGPU – Crédit Nvidia.....	190

Résumé

La problématique du transport de la donnée médicale, de surcroît nominative, comporte de nombreuses contraintes, qu'elles soient d'ordre technique, légale ou encore relationnelle. Les nouvelles technologies, issues particulièrement des grilles informatiques, permettent d'offrir une nouvelle approche au partage de l'information. En effet, le développement des intergiciels de grilles, notamment ceux issus du projet européen EGEE, ont permis d'ouvrir de nouvelles perspectives pour l'accès distribué aux données. Les principales contraintes d'un système de partage de données médicales, outre les besoins en termes de sécurité, proviennent de la façon de recueillir et d'accéder à l'information. En effet, la collecte, le déplacement, la concentration et la gestion de la donnée, se fait habituellement sur le modèle client-serveur traditionnel et se heurte à de nombreuses problématiques de propriété, de contrôle, de mise à jour, de disponibilité ou encore de dimensionnement des systèmes. La méthodologie proposée dans cette thèse utilise une autre philosophie dans la façon d'accéder à l'information. En utilisant toute la couche de contrôle d'accès et de sécurité des grilles informatiques, couplée aux méthodes d'authentification robuste des utilisateurs, un accès décentralisé aux données médicales est proposé. Ainsi, le principal avantage est de permettre aux fournisseurs de données de garder le contrôle sur leurs informations et ainsi de s'affranchir de la gestion des données médicales, le système étant capable d'aller directement chercher la donnée à la source.

L'utilisation de cette approche n'est cependant pas complètement transparente et tous les mécanismes d'identification des patients et de rapprochement d'identités (*data linkage*) doivent être complètement repensés et réécrits afin d'être compatibles avec un système distribué de gestion de bases de données. Le projet RSCA (Réseau Sentinelle Cancer Auvergne – www.e-sentinelle.org) constitue le cadre d'application de ce travail. Il a pour objectif de mutualiser les sources de données auvergnates sur le dépistage organisé des cancers du sein et du côlon. Les objectifs sont multiples : permettre, tout en respectant les lois en vigueur, d'échanger des données cancer entre acteurs médicaux et, dans un second temps, offrir un support à l'analyse statistique et épidémiologique.

Mots-clés : grille informatique, identification des patients, bases de données distribuées, dépistage des cancers

Abstract

Nominative medical data exchange is a growing challenge containing numerous technical, legislative or relationship barriers. New advanced technologies, in the particular field of grid computing, offer a new approach to handle medical data exchange. The development of the gLite grid middleware within the EGEE project opened new perspectives in distributed data access and database federation. The main requirements of a medical data exchange system, except the high level of security, come from the way to collect and provide data. The original client-server model of computing has many drawbacks regarding data ownership, updates, control, availability and scalability. The method described in this dissertation uses another philosophy in accessing medical data. Using the grid security layer and a robust user access authentication and control system, we build up a dedicated grid network able to federate distributed medical databases. In this way, data owners keep control over the data they produce.

This approach is therefore not totally straightforward, especially for patient identification and medical data linkage which is an open problem even in centralized medical systems. A new method is then proposed to handle these specific issues in a highly distributed environment. The *Sentinelle* project (RSCA) constitutes the applicative framework of this project in the field of cancer screening in French Auvergne region. The first objective is to allow anatomic pathology reports exchange between laboratories and screening structures compliant with pathologists' requirements and legal issues. Then, the second goal is to provide a framework for epidemiologists to access high quality medical data for statistical studies and global epidemiology.

Keywords : grid computing, patient identification, record linkage, distributed databases, cancer screening